ARMY RESEARCH LABORATORY

Proceedings of the Second Annual U.S. Army Conference on Applied Statistics, 23-25 October 1996

by Barry A. Bodt
EDITOR

Hosted by:
TEXCOM EXPERIMENTATION CENTER

Cosponsored by:
U.S. ARMY RESEARCH LABORATORY
U.S. MILITARY ACADEMY
U.S. ARMY RESEARCH OFFICE
WALTER REED ARMY INSTITUTE OF RESEARCH
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
TRADOC ANALYSIS CENTER-WSMR

ARL-SR-59 July 1997

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5067

ARL-SR-59 **July 1997**

Proceedings of the Second Annual U.S. Army Conference on Applied Statistics, 23–25 October 1996

Barry A. Bodt, Editor
Weapons and Materials Research Directorate, ARL

Hosted by:
TEXCOM EXPERIMENTATION CENTER

Cosponsored by:
U.S. ARMY RESEARCH LABORATORY
U.S. MILITARY ACADEMY
U.S. ARMY RESEARCH OFFICE
WALTER REED ARMY INSTITUTE OF RESEARCH
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
TRADOC ANALYSIS CENTER-WSMR

Abstract

The second U.S. Army Conference on Applied Statistics was held 23-25 October 1996 at the Monterey Beach Hotel, Monterey, CA, and hosted by the TEXCOM Experimentation Center at nearby Fort Hunter Liggett. The conference was cosponsored by the U.S. Army Research Laboratory; the U.S. Army Research Office; the U.S. Military Academy; the U.S. Army Training and Doctrine Command Analysis Center, White Sands Missile Range; the Walter Reed Army Institute of Research; and the National Institute for Standards and Technology. Papers given at the conference addressed the development of new statistical techniques, application of existing methodologies to Army problems, and panel discussion of statistical challenges in an Army setting. A special session was included to commemorate Fort Hunter Liggett, the dedicated civilians and military who have worked there, and the countless contributions to Army testing that were developed and practiced there. This document is a compilation of available papers offered at the conference.

FOREWORD

The second U.S. Army Conference on Applied Statistics was held 23–25 October 1996 at the Monterey Beach Hotel, Monterey, CA and hosted by the TEXCOM Experimentation Center (TEC) at nearby Fort Hunter Liggett. The conference was cosponsored by the U.S. Army Research Laboratory (ARL); the U.S. Army Research Office (ARO); the U.S. Military Academy (USMA); the Training and Doctrine Command (TRADOC) Analysis Center, White Sands Missile Range (WSMR); the Walter Reed Army Institute of Research (WRAIR); and the National Institute for Standards and Technology (NIST). The U.S. Army Conference on Applied Statistics is successor to the U.S. Army Conference on the Design of Experiments, an historic series of meetings that formally concluded in 1994 after forty years of service to the Army. Today's Army faces challenges that are far ranging and encompass many topics in which probability and statistics have a contribution to make, in addition to experimental design. This new conference reflects a broadening of scope with the goal to promote the practice of statistics in the solution of diverse Army problems.

The second conference continued in this new direction. Toward statistical education, the conference was preceded with a short course, "Quality Control: Modeling the Deming Paradigm," given by Prof. James R. Thompson of Rice University. Distinguished speakers from academia spoke during invited general sessions: Prof. C. R. Rao, Penn State University; Prof. Ulf Grenander, Brown University; and Prof. Rob Kass, Carnegie-Mellon University. A special session was included to commemorate Fort Hunter Liggett, the dedicated civilians and military who have worked there, and the countless contributions to Army testing that were developed and practiced there. As the program included will indicate, many prominent individuals participated in this special session. The conference was, however, especially pleased to welcome the Honorable Philip E. Coyle, Director, Defense Directorate of Operational Test and Evaluation (DOTE), Office of the Secretary of Defense (OSD); Mr. Walter W. Hollis, Deputy Undersecretary of the Army for Operations Research (DUSA-OR); Prof. Herman Chernoff, Harvard University; Dr. Ernest Seglie, Science Advisor, DOTE; and Dr. Marion Bryson, Former Director of the Combat Developments Experimentation Center (CDEC). The conference was completed with contributed sessions where talks developed new methodology, detailed successful applications, or requested guidance from a panel of experts in attacking an Army problem that had resisted standard statistical approaches.

The Executive Board for the conference recognizes Drs. Douglas Tang, WRAIR, and Mark Vangel, NIST, for assisting with conference details; Dr. Barry Bodt, ARL, for general conference administration and proceedings; and Dr. Carl Russell, TEC, for hosting the conference and handling all local arrangements. Special thanks is due to Mrs. Patricia Winters, TEC, who served as site coordinator for the conference.

Executive Board		
Robert Burge (WRAIR)	Barry Bodt (ARL)	Deloras Testerman (YPG)
Malcolm Taylor (ARL)	Eugene Dutoit (AIS)	Jerry Thomas (ARL)
Douglas Tang (WRAIR)	Jock Grynovicki (ARL)	Mark Vangel (NIST)
David Cruess (USUHS)	Carl Russell (TEC)	Paul Deason (TRAC-WSMR)

INTENTIONALLY LEFT BLANK.

Table of Contents*

* This Table of Contents contains only the papers that appear in the Proceedings.

FINAL PROGRAM
SECOND U.S. ARMY CONFERENCE ON APPLIED STATISTICS

21-25 October, 1996

Hosted by the TEXCOM Experimentation Center

Cosponsored by:

U.S. Army Research Laboratory
U.S. Army Research Office
National Institute of Standards and Technology
TRADOC Analysis Center—WSMR
United States Military Academy
Walter Reed Army Institute of Research

Monday, 21 October 1996

0800 - 0900 REGISTRATION (Monterey Beach Hotel)

0900 - 1200 TUTORIAL:

QUALITY CONTROL: MODELING THE DEMING PARADIGM
James R. Thompson, Rice University

1200 - 1330 Lunch

1330 - 1600 TUTORIAL

Tuesday, 22 October 1996

0800 - 1200 TUTORIAL

1200 - 1330 Lunch

1330 - 1600 TUTORIAL

Wednesday, 23 October 1996

0800 - 0900 REGISTRATION (Monterey Beach Hotel)

0900 - 0930 CALL TO ORDER:

Conference Chairman: Barry A. Bodt, U.S. Army Research Laboratory

OPENING REMARKS:

Conference Host: Carl Russell, Chief Scientist, TEXCOM Experimentation Center

0930 - 1200 GENERAL SESSION I

Chair: Robert L. Launer, ARO

0930 - 1030 KEYNOTE ADDRESS

PRE AND POST LEAST SQUARE: THE EMERGENCE OF ROBUST INFERENCE
C.R. Rao, Penn State University

1030 - 1100 Break

1100 - 1200 A PATTERN THEORETIC APPROACH TO ATR
Ulf Grenander, Brown University

1200 - 1330 Lunch (fajita buffet)

1330 - 1500 CONTRIBUTED SESSION I (in parallel with Clinical Session I)

Chair: Robert Burge, Walter Reed Army Institute of Research

PERFORMANCE-BASED METRICS FOR THE DIGITIZED BATTLEFIELD: PRAIRIE WARRIOR 96–MANEUVER COMMAND AND CONTROL SYSTEM
Jock O. Grynovicki*, Michael Golden, Kragg P. Kysor, and Dennis Leedom; U.S. Army Research Laboratory

THE IMPACT OF FIELD OF VIEW ON THE PERFORMANCE OF SOME INFANTRY TASKS
Eugene Dutoit, Dismounted Battlespace Battle Lab, Fort Benning

EMPIRICAL EVALUATION OF THE FACT EXCHANGE PROTOCOL, A TACTICAL NETWORK PROTOCOL
Ann E. M. Brodeen*, Maria C. Lopez, George W. Hartwig, Jr., and Michael J. Markowski; U.S. Army Research Laboratory

1330 - 1500 CLINICAL SESSION I (in parallel with Contributed Session I)

Chair: Bernard Harris, University of Wisconsin-Madison

Panel: Donald P. Gaver, U.S. Naval Postgraduate School
Robert L. Launer, U.S. Army Research Office

Wednesday, 23 October 1996 (Continued)

CLINICAL SESSION I, *cont.*

FRAMEWORK FOR EVALUATING THE VALIDITY OF OPERATIONAL TESTS AND FIELD EXPERIMENTS
Rick Kass, Test and Experimentation Command, Fort Hood

STATISTICAL ENHANCEMENTS AND VALIDATION OF A MIX MODEL METHODOLOGY
Bruce W. Gafner, TRADOC Analysis Center—WSMR

1500 - 1515 Break

1515 - 1715 CONTRIBUTED SESSION II (in parallel with Clinical Session II)

Chair: David Cruess, Uniformed Services University for the Health Sciences

ANALYSIS OF SYNTHETIC PROPORTIONS
Carl Russell, TEXCOM Experimentation Center

AN INTELLIGENT HIERARCHICAL DECISION ARCHITECTURE FOR OPERATIONAL TEST AND EVALUATION
MAJ Suzanne M. Beers*, Air Force Operational Test and Evaluation Center
George J. Vactsevanos, Georgia Institute of Technology

AN APPROACH TO GENERATING BAYESIAN PROBABILITY OF BELIEF IN MISSILE P_k
Michael B. Dewitz* and Paul M. Ellner, Army Materiel Systems Analysis Activity

PREDICTIVE QUALITY CONTROL CHARTS
MAJ David H. Olwell, United States Military Academy

1515 - 1715 CLINICAL SESSION II (in parallel with Contributed Session II)

Chair: W. J. Conover, Texas Tech University

Panel: C.R. Rao, Penn State University
Edward J. Wegman, George Mason University

PERMUTATION-BASED EXTRAPOLATED REGRESSION ESTIMATES
David W. Webb, Army Research Laboratory

POWER SIMULATION OF THE WILCOXON SIGNED-RANK TEST
Thomas R. Walker, Aberdeen Test Center

SELECTION OF APPROPRIATE TECHNIQUES FOR CLASSIFICATION OF PARTS
Barnard H. Bissinger, Navy Inventory Control Point (NAVICP)

Thursday, 24 October 1996

Forty Years of Experimentation at Fort Hunter Liggett

0800 - 0900 GENERAL SESSION II

Chair: Ernest Seglie, Science Advisor, Operational Test and Evaluation (DOTE)

IMPROVING USE OF STATISTICS IN ARMY TEST AND EVALUATION
Herman Chernoff, Harvard University

0900 - 0915 Break

0915 - 1045 SPECIAL SESSION I

Experimentation at Fort Hunter Liggett, Overview and Early Years

Chair: James Prouty, Commander, TEXCOM Experimentation Center

OVERVIEW OF EXPERIMENTATION AT FORT HUNTER LIGGETT
Carl Russell, Chief Scientist, TEXCOM Experimentation Center

RECOLLECTIONS OF THE FIRST YEARS OF CDEC: SEPTEMBER 1956 TO JUNE 1958
Floyd Hill, Former Associate Director, Research Office, Evaluation Command

INSTRUMENTATION TO MEET EARLY CDEC DATA NEEDS
Henry Alberts, Former Instrumentation Chief, SRI Research Office

1045 - 1100 Break

1100 - 1200 GENERAL SESSION III

Chair: Marion Bryson, Former Director of CDEC

FROM STOPWATCH TO COMPUTER: THE TRANSITION OF EXPERIMENTATION AT FORT HUNTER LIGGETT
Walter Hollis, Deputy Undersecretary of the Army for Operations Research

1200 - 1315 Lunch (cold cut buffet)

1315 - 1415 GENERAL SESSION IV

Chair: Walter Hollis, Deputy Undersecretary of the Army for Operations Research

EVOLUTION OF EXPERIMENTATION AT FORT HUNTER LIGGETT IN THE 1970s AND 1980s
Marion Bryson, Former Director of CDEC

1415 - 1430 Break

Thursday, 24 October 1996 (Continued)

Forty Years of Experimentation at Fort Hunter Liggett

1430 - 1600 SPECIAL SESSION II

Experimentation at Fort Hunter Liggett: Special Aspects

Chair: Bill West, Former Chief Scientist, TEXCOM Experimentation Center

SOME CLASSIC DATA FROM FORT HUNTER LIGGETT—THEIR CONTINUED RELEVANCE TODAY
Brian Barr, TEXCOM Technical Director

THE IMPORTANCE OF HIGH-RESOLUTION RTCA IN OPERATIONAL TESTING
Ernest Seglie, Science Advisor, Operational Test and Evaluation (DOTE)

FORT HUNTER LIGGETT ON WHEELS—MOBILE INSTRUMENTATION FOR COMBAT FIELD EXPERIMENTS
Ed Buntz and Mike Tedeschi
Instrumentation Division, TEXCOM Experimentation Center

1600 - 1615 Break

1615 - 1715 SPECIAL SESSION III

Panel Discussion on the Future of Field Experimentation

Moderator: Brian Barr, TEXCOM Technical Director

Panelists:

Dr. Ernest Seglie
Science Advisor, Operational Test and Evaluation (DOTE)

Mr. Walter W. Hollis
Deputy Undersecretary of the Army for Operations Research

Dr. Marion R. Bryson
Former Director of CDEC

1830 - Banquet, Monterey Beach Hotel

Speaker: Honorable Philip E. Coyle III
Director, Operational Test and Evaluation, OSD

Friday, 25 October 1996

0800 - 0930 CONTRIBUTED SESSION III

Chair: Deloris Testerman, Yuma Proving Ground

MODELING LOSS EXCHANGE RATIOS AS INVERSE GAUSSIAN VARIATES: IMPLICATIONS
MAJ David H. Olwell, United States Military Academy

CASTFOREM VERIFICATION AND VALIDATION PROCESS
Douglas C. Mackey, TRADOC Analysis Center—WSMR

INVESTIGATION INTO VARYING ARTILLERY COMMUNICATION TIMES IN CASTFOREM
Paul J. Deason, TRADOC Analysis Center—WSMR

0930 - 0945 Break

0945 - 1115 CONTRIBUTED SESSION IV

Chair: Eugene Dutoit, Battle Space Battle Lab, Fort Benning

NONPARAMETRIC DENSITY ESTIMATION BY QUADRATIC OPTIMIZATION
Joseph Collins, Army Research Laboratory

MODELING THE EFFECTS OF RECOIL OF SHOULDER-FIRED WEAPONS
Jock O. Grynovicki*, William Harper, Kathy Leiter, Sam Ortega, and Kragg Kysor; U.S. Army Research Laboratory

PROJECTION METHODS FOR GENERATING MIXED-LEVEL FRACTIONAL FACTORIAL DESIGNS
Alonzo Church, Jr., Church Associates, Inc.

1115 - 1130 Break

1130 - 1230 GENERAL SESSION V

Chair: Barry A. Bodt, U.S. Army Research Laboratory

SOME EXAMPLES OF BAYESIAN INFERENCE
Rob Kass, Carnegie-Mellon University

1230 ADJOURN

Performance-Based Metrics to Assess Battlefield Visualization:
Prairie Warrior 96 -- Maneuver Command and Control System (MCS/P)

Dr. Jock O. Grynovicki *Army Research Laboratory-Hum. Res. & Engr. Dir.*
Mr. Michael Golden *Army Research Laboratory-Hum. Res. & Engr. Dir.*
Mr. Kragg Kysor *Army Research Laboratory-Hum. Res. & Engr. Dir.*
Dr. Dennis Leedom *Army Research Laboratory-Hum. Res. & Engr. Dir.*

Abstract

One of the U.S. Army Research Laboratory's (ARL's) Science and Technology Objective (STO) research projects is to develop standardized field-operational soldier performance metrics to quantify integrated soldier-information system performance on the digital battlefield. This research effort is intended to help the Army leadership assess the impact of digitization on individual soldier and staff performance. These measurement scales directly support the Joint Venture Axis Five and Seven and Rolling baseline assessment of digital information system technology during Advanced Technology Demonstrations, Advanced Warfighting Experiments, and related Force XXI and Army-After-Next field activities.

In conjunction with this project, ARL supported the Battle Command Battle Laboratory (BCBL) and the TRADOC Analysis Center (TRAC) in studying Battlefield Visualization issues during the Prairie Warrior 96 exercise (PW 96). Specifically, ARL's emphasis was on the Maneuver Control System/Phoenix (MCS/P) beta Battlefield Operation Systems (BOS) software that was designed to enhance the Mobile Strike Force (MSF) soldier and staff performance during the exercise by providing a clear understanding of the current state of a battlefield situation with relation to the enemy and environment.

The paper specifically describes efforts to define and measure soldier MCS/P information interface functionality and usability. The report includes lessons learned from PW 96 and describes how the evaluation methods and metrics were developed and improved to produce an evaluation package that can be use in other Advanced Warfighting Experiments (AWEs), Command Post Exercises (CPXs), and simulation exercises. Results of the behaviorally anchored rating scale and usability index administered to the MSF during PW 96 are presented.

Key Words Prairie Warrior 96, MCS/P, performance metrics, behavior anchor scales, soldier system interface

1 Introduction

The U.S. Army Research Laboratory (ARL) supported the Battle Command Battle Laboratory and the TRADOC Analysis Center (TRAC) in studying Battlefield Visualization issues during the Prairie Warrior 96 exercise (PW 96). Specifically, ARL's emphasis was on the Maneuver Control System/Phoenix (MCS/P) beta Battlefield Operation Systems (BOS) software that was designed to enhance the Mobile Strike Force (MSF) soldier and staff performance during the exercise by providing a clear understanding of the current state of a battlefield situation with relation to the enemy and environment. The soldier and staff MCS/P interface was assessed by ARL through

the administration of anchored rating scale questionnaires to the MSF participants and observations during SIMEX I and PW 96. In this study, we measured digital effects in terms of attitude change, behavior change, command staff task performance, and soldier-computer interface effectiveness.

Specifically, quantitative psychometric methods were used in the development of behaviorally anchored rating scales and standardized task performance metrics to evaluate integrated staff and soldier information system interface performance on MCS/P. The rating scale methodology used a five-point Likert-type scale to quantify MCS/P functionality. These metrics addressed critical functional dimensions of staff performance within the Deliberate Decision Making Process that included: (1) Mission Analysis (2) Course of Action (COA), (3) Information Assimilation, (4) Generation of Messages and Reports, (5) Workload Distribution and (6) Development, Distribution and Maintenance of Situation Awareness.

To study and improve soldier-computer interface software design, a heuristic evaluation was administered to the MSF. This evaluation used a usability index developed by ARL for measurement that focused on important soldier MCS/P interface design issues involving such characteristics as speed, utility, flexibility, consistency, intuitiveness, feedback, demand on memory, error recovery, and fatigue. These principles are based on human-system interface research outlined by Molich and Nielsen (1990). This index also used a five-point Likert-type scale. The paper specifically describes efforts to define and measure soldier MCS/P information interface functionality and usability. The report includes lessons learned from PW 96 and describes how the evaluation methods and metrics were developed and improved to produce an evaluation package that can be transition for use in other Advanced Warfighting Experiments (AWEs), Command Post Exercises (CPXs), and simulation exercises. Results of the behaviorally anchored rating scale and usability index administered to the MSF during PW 96 are presented.

2 Prairie Warrior

Command and General Staff College(CGSC) designed the exercise to provide the Command and General Staff Officers Course (CGSOC) students with an experience similar to a Warfighter and provide an opportunity to execute decision-making processes. Operations in a joint and multinational environment were simulated.

The Command and General Staff Officers course A308, Battle Command Elective, provided the staff and systems training for the MSF. This included classroom instruction in MSF concepts and tactics, techniques, and procedures (TTP); hands-on training for MCS/P; two simulation exercises (SIMEXes) and the final exercise Prairie Warrior 96 (PW 96).

As stated in the PW 96 Final Report (1996), "principal units (located at Fort Leavenworth, Kansas, unless otherwise noted) included a Combined Joint Task Force (CJTF); Combined Forces Component Commanders; a Theater Support Command (TSC), represented by the 310th Theater Army Area Command (TAACOM) operating from Fort Lee, Virginia; a student-led corps and subordinate U.S. and multinational divisions; a student led MSF; a student-led Marine Air Ground Task Force; Analysis and Control Elements (ACEs) staffed by Military Intelligence Officer Advanced Course (MI OAC) students at Fort Huachuca, Arizona; Analysis and Control Teams (ACTs) staffed by MI Officer Basic Course (OBC) students; and a Synchronization Cell,

operating from Maxwell Air Force Base, Alabama." The MSF used advanced systems with potential 2010 technology.

2.1 Maneuver Control System/Phoenix (MCS/P) (beta).

The MCS/P (beta) was the central digitized platform used in PW 96. It was a prototype computerized battle command system. This system provided a common picture of the battlefield overlaid on Defense Mapping Agency (DMA) digital maps. There was capability inherent in the system to synchronize the battle plan based on the assessment or presentation of near-real-time information and assessments from staff and subordinate commanders. MCS/(P) had the capabilities of conveying current information about location, strength, and other pertinent information for both friendly and enemy forces. A total of 56 MCS/Ps were used in PW 96 by the MSF, 25 of them in the NSC and the remainder in the Leadership Development Center (LCD). This command and control system had the following capabilities:

1. Receive enemy and friendly feeds
2. Build and manipulate databases
3. Generate and display reports
4. Create Situational Awareness
5. Build overlays
6. Operate planning
7. Wargaming
8. Send & receive information and briefs

The distribution of MCS/Ps and other systems in the MSF is shown in Figure 1.

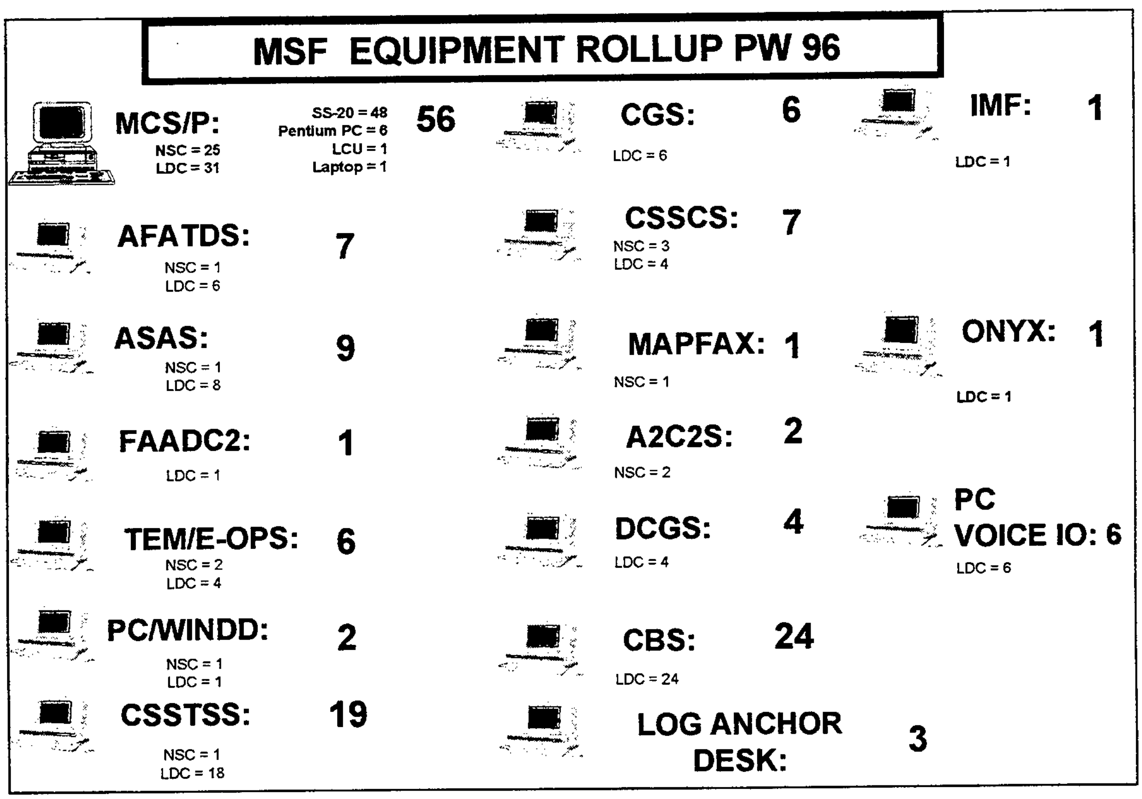


Figure 1: Battlefield Operating Systems Used in PW 96

Seventeen systems were included in the PW 96 Army Tactical Command and Control Systems (ATCCS). They included: (1) Maneuver Control System/Phoenix (MCS/P), (2) Common Ground Sensor (CGS), (3) Intelligent Mine Field (IMF), (4) Advanced Field Artillery Tactical Data System (AFATDS), (5) Combat Service Support Control System (CSSCS), (6) All Source Analysis System (ASAS), (7) Map Fax System (MAPFAX), (8) Onyx Graphics System (ONYX), (9) Forward Area Air Defense Command, Control, and Intelligence (FAAD C2), (10) Army Airborne Command and Control System (A2C2S), (11) Terrain Evaluation Module-Engineer Operations System (TEM/E-OPS), (12) Downsized Ground Control Station (DGCS), (13) Voice Activation (PC Voice IO), (14) Windows Desktop Display (WINDD), (15) Corps Battle Simulation (CBS), (16) Combat Service Support Training support Simulation (CSSTSS) and (17) Knowledge Based Logistics Planning Shell (KBLPS/ Log Anchor Desk).

2.2 Digital Network Architecture.

Both the National Simulation Center and the Leadership Development Center contained elements of the MSF and the Bell Hall operation contained the II Corps. The buildings were interconnected with high capacity data lines and interconnected with data to simulate tactical data communications (see Figure 2).

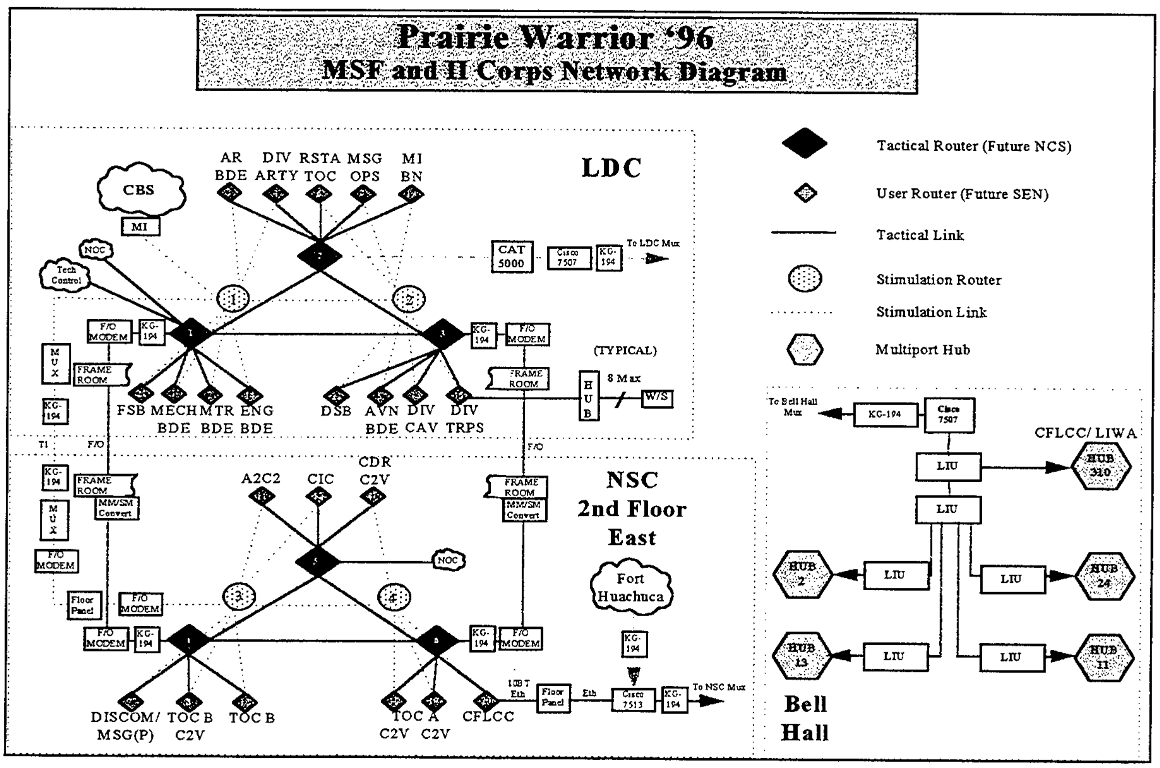


Figure 2: MSF and II Corps Element Network

2.3 Overall Staff Organization

The commander exercised command and control of the MSF through two distinct and separate tactical operations centers (TOCs A and B), the Combat Information Center (CIC), and the Analysis and Control Element (ACE). Additionally, an ad hoc TOC charged with reconnaissance, intelligence, surveillance, and target acquisition (RISTA) tasks evolved from the military intelligence battalion headquarters. TOC A focused on current operations and was employed forward on the battlefield, while TOC B focused on future or subsequent operations and deployed in the rear. The ACE and the RISTA TOC provided a mix of collection and analysis. The CIC managed information to support the commander. Further, a mock-up of the Army Airborne Command and Control System (A2C2S) provided the MSF commander a platform to facilitate moving around the battlespace.

TRADOC's concept of the CIC was defined by the BCBL as "the means to meet the information needs of the commander, staff, and subordinate units. As such, the CIC was used to gather, integrate, and synthesize information and/or information products into a focused, division-level database for the commander and the tactical operations centers." The CIC focused information searches to support information requirements, based on Commander's Critical Information Requirements (CCIR), and processed information into an integrated, coherent product called the Relevant Common Picture (RCP). The concept included requirements for the CIC to meet specific requests for information not contained in the RCP that the commander and staff may have required or may later require.

An Assistant Chief of Staff for the CIC provided direct supervision of the seven man CIC operation that included (1) Fusion/OPS, (2) Maneuver, (3) Engineer, (4) Intelligence, (5) Air Defense, (6) Combat Service Support, and (7) Field Artillery. A staff officer manned each of the separate MCS/P work stations within the CIC and represented each of the battlefield operating systems (BOS). A Fusion and OPS workstation (Command and Control functional area) served as the integration station for information provided by the other members of the CIC. In addition to the workstations, the CIC had a large screen display (60 inch monitor) connected to the Fusion/OPS Workstation.

The functional workstations used information transmitted from BOS workstations throughout the MSF (e.g., MCS/P, TEM-E/OPS, AFATDS) depending on the functions performed. Additionally, the CIC monitored a "brick" radio tuned to the MSF operations and intelligence (O & I) radio net. This voice network provided the opportunity for noting updates of the current operations. This "stovepipe" information flow into the CIC from the appropriate areas or subordinate units provided the functional workstation the capability to manipulate or refine data to provide the required information. Once prepared, the functional workstations transmitted the information to Fusion/OPS which condensed, verified, and distributed the information to lower echelons.

3 Performance-Based Metrics Methodology

With Task Force XXI and the Army After Next initiatives, the Army has initiated a campaign to evaluate Advanced Warfighting Experiments that will leverage superior technology to build the Army of tomorrow. The central and essential feature of this Army will be its ability to exploit information, which will lead to quick and decisive victory. Soldiers will be the most important element of Force XXI, for it is through quality soldiers that the full power of technology will be realized.

ARL assisted the TRAC in assessing soldier information system interface and staff coordination and performance using MCS/P. The administration of anchor scale surveys, direct observation, and video tape recording of MCS/P during SIMEX I and PW 96 provided the information base for this assessment. This research used psychometric principles, staff coordination behavior based methodology (Leedom & Simon, 1995) and Human System Interface research (Molich & Nielsen, 1990) to develop the anchor scales and standardized task performance metrics to evaluate integrated staff and soldier information system performance on the MCS/P BOS.

The Universal Joint Task List (UJTL) was used to identify essential tasks that a combat commander is required to perform in exercising command and control. This list serves as an interoperability tool to help commanders construct their joint mission essential task list. It is a comprehensive hierarchical listing of the tasks that can be performed by a joint military force. UJTL is organized into four separate parts by the level of war: Strategic level - National military tasks, Strategic level - Theater tasks, Operational level, and Tactical level tasks. Each task in the UJTL is individually indexed to reflect its placement in the structure. Thus, the UJTL provides a standard reference system for users to address and report requirements, capabilities, or issues and as such formed the Command Staff task baseline around which ARL developed its standardized soldier performance metrics research efforts.

3.1 Behavior-Anchored MCS/P Function Support Assessment

Two behavior-anchored scale assessment instruments were developed and administered to the MSF during SIMEX I and PW 96. Utilizing the decision level UJTL tasks as a foundation, the first instrument focused on the interrelationship between the Division Staff functions or processes and MCS/P. ARLs' metrics development methodology established a crosswalk of FM 101-5 Deliberate Decision Making Processes(DDMP) with the MCS/P software modules believed to support critical (identified) command, staff task execution. FM 101-5 states that a staff supports the science of control in four primary ways: (1) gathers and provide information to the commander, (2) makes estimates of the set of actions required, (3) prepares plans and orders, and (4) measures organization behavior. To perform this type of support, the Division Staff and commanders use the DDMP which requires staff coordination between and within echelons. It can be assumed that the MCS/P battle command system capabilities were developed to support these processes. Figure 3 depicts the ARL crosswalk of MCS/P beta software capabilities with the key staff tasks of the DDMP.

Given the Crosswalk matrix, ARL-HRED developed six (6) key behavior-based staff coordination evaluation dimensions to assess the ability of the digitized maneuver command control system MCS/P to support the DDMP and staff coordination. These six dimensions are listed in Table 1. Each dimension, is defined in terms of sub-dimensions and specific, operationally relevant, staff related behavior. The behavior anchor scale format standardized the perception of the MSF as to what each dimension was trying to assess and experimentally reduced the response variability. Definitions and descriptions for MCS/P supporting the type of behavior for greatly facilitated, borderline, and greatly hindered performance were developed to include example task. The written descriptions of the levels of performance for each sub-dimension were assigned values of 5 through 1 (one reflecting that it hindered performance, three being the same as manual methods, and five that it facilitated performance) to serve as anchors for the five-point Likert type scale. Guidelines

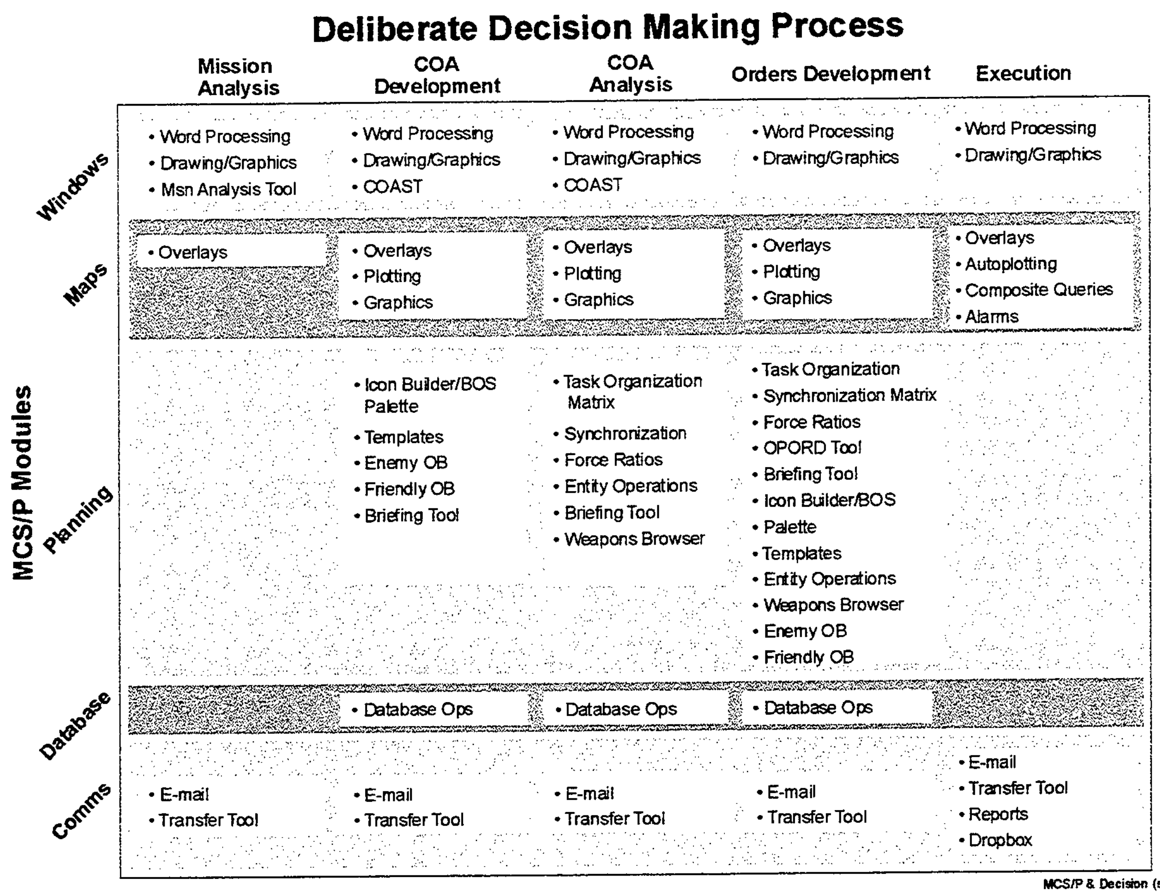


FIGURE 3: Crosswalk of MCS/P Capabilities vs. Key Deliberate Decision Making Processes. OB: Order of Battle; OPORD: Operation Order; COAST: Course of Action Situational Template

were prepared to assist the MSF in assessing how well the staff performed. The MSF assessed the key Maneuver Staff functions after SIMEX 1, which was used as a training exercise, and PW 96, the main combat exercise.

3.2 Task-Centered Usability Assessment

The second standardized instrument developed by ARL-HRED, focused on the usability of the individual soldier-MCS/P system interface. Certain system design characteristics have been defined in the literature that reflect platforms with good interface usability (Nielsen & Molich, 1990). These design characteristics were used by ARL-HRED to focus on rating 12 staff tasks on sixteen interface usability and graphics issues as shown in Table 2. These characteristics include whether the computer system contains simple and natural dialogue, reflects doctrine or "speaks the user language," minimizes user memory load, remains consistent between different modules, provides feedback, provides clearly marked exits, provides shortcuts, and prevents errors.

The usability factor has a direct impact on the tactical decision-making process. Malfunctions in system usability lead to underlying error patterns such as attention fatigue, excessive mental workload, inappropriate priorities, delays in tempo, and ultimately, communication failures. These error problems can lead to more serious tactical failures such as inadequate battle plans,

Table 1: Behavior Evaluation Dimensions

Evaluation Dimension	Sub Dimension	Behavior Anchor Focus
Mission Planning and Refinement	Impact on Mission Analysis	Automated information being readily available and assessable to facilitate horizontal and parallel planning
Mission Planning and Refinement	Impact on COA Development	Coordinated input into the developing COAs of key staff perspectives
Mission Planning and Refinement	Impact on COA Analysis	Staff simultaneously analyzing alternative COAs by maintaining a shared common understanding of mission intent, joint identification of COA problems, branch contingencies, etc.
Information Assimilation	Assimilation of digitized messages	Finding, reviewing, and assimilating information from text messages to obtain CCIR
Information Assimilation	Assimilation of digitized graphics	Finding, reviewing, and assimilating information from graphical display to obtain CCIR
Generation of Messages and Reports	Enhance ability to prepare orders and reports	Supporting the staffs' ability to prepare and send desired messages and reports
Situational Awareness	Real-time asses to data sources at all echelons for effective CCIR-based push/pulls?	Staff maintaining a shared, real-time awareness of the battlespace which is formulated into a coordinated RCP. Selective filtering and assimilation of situation-based information.
Situational Awareness	Facilitate effective monitoring of critical events and receipt of critical messages	How digitization assisted the battle staff in keeping each element aware and informed of critical events and factors.
The Relevant Common Picture	Facilitate development and maintenance of a coordinated relevant common picture?	The formulation of the RCP graphic visualizations and initial information dissemination. Staff automatic situation information monitoring. Automated graphic aids for timely RCP and follow-on distribution?
The Relevant Common Picture	Facilitate distribution of the relevant common picture updates to all battle command elements?	Timely distribution of the RCP graphic visualizations and information updates. Automated situation monitoring. Automated graphic aids for timely RCP updating and follow-on distribution?
Workload Distribution	Appropriately distribute mission tasks between staff	Mission task prioritization and workload distribution.

inadequate reporting, fratricide, lack of coordination, and inadequate situational awareness. Understanding the individual soldier-MCS/P system interface also signals the movement to correct lapses and underlying error problems with the system interface, and in turn prevent major system failures and significantly increase the chance of success in combat operations on the battlefield. In application, a heuristic evaluation was done by having the MSF rate all of the tasks for each issue on a scale of one to five (one being the worst and five the best) after having used MCS/P during SIMEX I and the actual Prairie Warrior exercise. Heuristic evaluation, as described by Nielsen and Molich 1990, is a method of usability analysis where a number of users are presented with an interface design and then expected to comment on it. As in the first instrument assessment tool, the soldiers' perception of the usability issues as related to the Maneuver Staff tasks was standardized using anchor scale methodology.

3.3 Performance-Based Metrics Participants

The information interface performance metrics were administered to the entire MSF during SIMEX I and PW 96. This force was primarily composed of students (88 Majors) from the Command and General Staff Officers Course A308 (CGSOC), Battle Command Elective (January-May 1996). This course provided the training for commanders and staff officers which included classroom instruction in MSF concepts and tactics, techniques, and procedures (TTP) and hands-on training for MCS/P. The class also included two simulation exercises (SIMEXes) and the culminating CGSOC exercise PW 96.

Table 2: Usability Index Issues and Maneuver Staff Tasks

Maneuver Staff Task	Usability Issues
Displaying & Manipulating Maps Plotting & Manipulating Units Building Overlays Templates Creating, Editing Updating Data Bases Building Friendly & Enemy Order of Battle	Tempo Utility Flexibility in use Prevent Fatigue Mirror Doctrine
Building & Modifying Synchronization Matrix	Provide process Short Cuts
Preparing Task Organizations	Consistency between Modules
Computing Force Ratios	Minimize demand on Memory
Preparing Briefings	Provide Feedback
Preparing Operation Orders	Good Error Recovery
Building & Displaying Alarms	Process Shortcuts
Sending & Receiving Inf.	Intuitiveness

4 Results & Discussion

4.1 MSF Responses

ARL administered the instruments to the entire MSF immediately after SIMEX 1 and the PW 96 exercise. A total of 84 (95 %) surveys were completed and returned after SIMEX I and 44 (54 %) after PW 96. Considering that the end of the PW 96 exercise coincided with graduation ceremonies and the students being assigned to other duty stations and packing to leave Fort Leavenworth, Kansas, a decline in the responses was not surprising. In addition, 8 students who were assigned to the MSF Commanders staff, did not use MCS/P during the PW 96 exercise and submitted blank questionnaires after PW 96. Since they did not use MCS/P since SIMEX I, they stated that their responses after PW 96 would be exactly the same as their response after the simulation exercise.

To address response bias, the SIMEX I response distribution of the 44 students that responded to both SIMEX 1 and PW 96 was compared to the response distributions of the 40 students that responded after SIMEX I but did not respond after PW 96. A non-parametric Chi-Square statistic was used to determine if the frequency response across the rating scale was statistically different between the two response groups for each question. In 93 % of all questions, no significant difference between the groups could be determined at the .05 significance level. By chance alone,

one would expect between two and three questions to be significant when testing at the .05 level. Thus, there is no evidence of a significant response bias.

4.2 Database Building, Manipulating, and Editing

The MCS/P was extremely flexible, useful, and reduced the time it took for the user to manipulate databases. More than 70% of the MSF respondents (Chi Square = 22.4, $p < .05$) liked the utility and speed of the MCS/P database capabilities, indicated that it mirrored doctrine, and was not as fatiguing as standard methods. The database system allowed the user to construct a list of friendly and enemy databases used most frequently by each staff element (G2, G3, engineer, etc.). However, only one record in a database could be located at a time. This serial editing and retrieval of information violated cognitive congruency between soldiers' expectations and MCS/P which caused the DDMP to break down and force the decision makers to invoke other cognitive tactics such as decision forestalling and assumption-based reasoning. Cognitive Congruency relates to the degree to which the information management and display paradigms are matched with the training and experience of the human operator. Because a database could contain more than 100 records, this record-locating procedure was time-consuming and resulted in an increased user workload and demand on memory. The majority of the respondents (56 %) felt that the error recovery was poor with only 10 % expressing an opinion that it was good (Chi Square = 19.3, $p< .05$). Regarding editing, the editor window was extremely cluttered and its use was neither intuitive nor consistent. One example of this confusion was the "edit records" procedure which used ADD, MODIFY, and DELETE commands. Another example was the "retrieve records" procedure which used the FETCH or QUERY command. The editor window should be simplified and the capability to handle more than one record at a time should be included in future MCS/P development.

4.3 Creating Situation Awareness

The usefulness of the MCS/P in creating situation awareness varied across the MSF echelons and was a function of information timeliness, accuracy, and detail. At the division level, the size and resolution of the 13-inch display prevented the commander from obtaining a detailed view of the entire MSF battlespace to visualize unit movement in a large area (275 k X 275 k). When this wide area battlespace was attempted to be viewed on the 13-inch computer monitor, the multitude of symbols and icons involved presented a cluttered display. This area of view limitation and display clutter (lack of Cognitive Congruency) did not fit the experienced based mental model of the commander and limited his insight. To deal with is limitation, division commanders relied on the map sheets to do mission planning and analysis. Fewer than 20 % of the respondents (Chi Square = 60.1 $p< .01$) felt that digitization facilitated mission analysis because of this limitation of battlefield visualization and stated in their comments that they relied on the map sheets to conduct mission planning.

The MCS/P, however, was a good tool for displaying information and allowing the staff elements (soldier), at the smaller brigade area of interest, to integrate this information to create situational awareness of their battlespace and conduct Course of Action (COA) development. More than 57% of the MSF rated the MCS/P as facilitating or greatly facilitating and 97 % felt that it supported the staffs' ability to monitor critical events and receive critical messages in a timely fashion as well as distribute the Relevant Common Picture (REP) and keep all elements aware of critical

situational change (Chi Square = 43.8, $p < .01$). The MSF user displayed maps, set features of the maps, plotted military units, and manipulated unit icons easily. The users were also able to easily zoom into an area of interest using either a raster graphics or vector map. The location tracker assisted the commander in locating positions on a map and cross-validate positions received from ASAS. The majority of the MSF participants (69 %) reported that the map tools mirror doctrine, were consistent (80 %) and had a moderate to low demand on memory (77 %). The armor brigade commander used the field-of-view tool to tactically position his units. Map features were easily displayed and the distance tracker tool was observed being used by the operations officer and engineer. The ability to display a unit's strength was a valuable tool in developing courses of action (COAs). The armor brigade commander in the MSF frequently used this tool to determine the effectiveness of the course of action that he had previously chosen.

At STARTEX of PW 96, there was a problem with data transfer as the MSF staff could not retrieve a database and transfer information automatically. To compensate for this deficiency, templates were created manually by drawing phase lines and units. This delayed the development of the RCP and diminished its accuracy regarding enemy positions. During SIMEX I, II, and PW 96, data flows to the CIC improved, the time it took to develop the current RCP decreased by 30 minutes on the average, but enemy location data were still one hour old and therefore not timely. In conclusion, the MCS/P has numerous tools to create the RCP, but when the information is neither timely nor accurate, relevant situational awareness will not be achieved. This deficiency degrades performance, especially during close battle operations.

4.4 Templates and Drawing Tools

The overlay building capability of MCS/P was the most utilized tool in MCS/P. The entire MSF was observed using and developing templates. Overlays provided what many staff members considered to be "snap shots" of the battlefield, which effectively flowed among the MSF. After PW 96, almost 70% of the MSF rated this capability as better or much better compared to standard manual methods (paper maps, grease pencils) (Chi Square = 15.1, $p< .01$). The MCS/P overlays facilitated planning and COA development (70% of the MSF responded that it supported or facilitated COA development) by providing brigade staffs the ability to "call up" adjacent units' overlays, as well as operation orders using Microsoft Word (WINDD). This overlay retrieval capability allowed the brigade staffs to review the MSF division branch plans as they were performing their missions and allowed the brigades to quickly develop FRAGOs. This tool was extremely flexible. There are various layers in which different items on a map are saved so they could be tailored for different echelons and functions. For example, the default layer was used to present shapes, lines, and text. The command layer contained units from a database, map features, and obstacles. The grid layer contained grid information. However, the drawing tool was limited and not user-friendly. The MCS/P during PW 96 allowed the user to draw line segments, polygons, ellipses, rectangles, and circles. The MSF spent inordinate amounts of time drawing arrows and phase lines and positioning icons on the map. By the end of PW 96, the MSF was much more proficient in developing their overlays. This is reflective in a significant shift in the ARL's survey rating distribution of the MCS/P to facilitate monitoring of critical events, receipt of critical messages, and develop, update, and distribute the RCP between SIMEX I and PW 96.(Chi-Square = 10.84, $p< 0.02$). This drawing tool needs to be upgraded by automating the placement of phase lines, arrows and other icons and figures. Building a template was complicated, involving a multitude of windows and menus. For example, it took seven commands to plot one symbol from an existing palette. A simple drawing and symbol menu needs to be developed. The soldier could

click on a desired symbol and then insert it at the position of his or her cursor as can be done using Microsoft Word {symbol 211 \f "Symbol" \s 11}.

4.5 Soldier-Computer Usability

The primary soldier MCS/P (Beta) interface deficiencies concerned stability, error recovery, simplicity, user feedback, and process shortcuts. The soldier had poor feedback (lack of knowledge of results) to know that the MCS/P was processing a function, such as retrieving a map, or when a system error occurs. This resulted in the soldier trying alternate sequences of button pressing to achieve the desired function. This resulted in further deterioration of the computer system, the system locking up, and the loss of previous work. Over a six hour period during PW 96, 3 instances of MCS/Ps locking up were observed. Future systems should provide a visual icon that shows the system is in the midst of processing, understandable error messages, and automatic system backup.

The need for improved system error recovery was reflected in the MSF's response to the Task Centered Usability Index regarding error recovery. The majority of the MSF that responded felt that the error recovery capability of MCS/P needed improvement. Less than 11% of the responses rated the MCS/P as having good or excellent error recovery capabilities (Chi Square = 10.7, $p < .02$).

Observation revealed that there was a need for consistency between various system functions. For example, the user could select QUIT, END, or EXIT to end different functions. There were too many menus and steps and no well-defined shortcuts to perform or quit a function.

The survey also revealed that, in general, the system increased the tempo of activity, was flexible, mirrored doctrine, and was not fatiguing to use more than ten hours with high and low periods of battlefield operations for various maneuver command and control tasks while in a stationary environment. Almost 60 % of the MSF felt that digitization (MCS/P) reduced the time it took for the battle staff to complete its tasks (Chi Square = 72.0, $p<.01$). Across all modules and staff tasks, the opinion of the MSF was that the system mirrored doctrine. Over 60% of the responses rated the MCS/P software, automated processes, and sequences as accurately or very accurately mirrored doctrine. Seventy nine percent of the MSFs' responses reflected that the fatique level experienced by the MSF was the same or less as performing the tasks manually (Chi Square = 20.9, $p< 0.1$).

4.6 Sending and Receiving Information

The e-mail function in MCS/P was extremely useful in maneuver command and control and the "transfer tool" was an outstanding software application within e-mail. This function allowed the soldier to send and receive messages, overlays, and reports quickly to and from selected locations on the battlefield. Seventy percent of the MSF felt that the MCS/P supported or enhanced the ability of the staff to distribute the RCP and keep all MSF elements aware of critical situational change (Chi Square = 15.1, $p<.01$).

The transfer tool window consisted of three columns split into upper and lower sections. The upper-left column listed overlays, the upper-middle listed databases, and the upper-right listed site addresses for potential recipients. As the user selected the overlays or databases to be sent and the

intended recipients, the overlays and databases then appeared in the bottom sections of their corresponding columns. This organized visual feedback allowed easy and quick selection and transfer of information.

The e-mail application also allowed the user to transfer overlays between and within echelons. To retrieve an overlay from another machine, the user had to know the other machine's address. The e-mail system was also extremely flexible in editing destination addresses, and receiving brigade and below command and control reports. This tool should remain as a standard feature of MCS/P.

4.7 Collaborative Staff Information System Interface

The MCS/P digital information technology offered significant improvements in MSF collaborative staff performance over current manual, paper-based, voice-communicated command staff products. The MSF staff, organized around the conceptual staff element (CIC), used the MCS/P system extensively for distribution and maintenance of situational awareness through continuous generation and distribution of RCP graphics and "Post-it Notes."

During PW 96, the MCS/P aided the staff's collaborative planning and their execution tasks greatly improved from the initial attempts of SIMEX I. However, some performance shortcomings remained unchanged throughout the PW 96 exercise. This was mainly from a combination of ineffective staff training on MCS/P, as well as MCS/P software deficiencies . The training problem included (1) the lack of command staff /CIC staff training as a collaborative team and (2) the lack of effective individual training on MCS/P functions, which resulted in the staff members spending too much time trying to decide how to execute a MCS/P function rather than spending time performing critical staff functions with MCS/P. The lack of an effective MCS/P user's manual greatly contributed to the training shortfall. The overall MSF staff skill levels on MCS/P software functionality continued to improve throughout PW 96, but as a unit, they never reached fully effective levels. Some individual users, possessing superior computer skills and expertise with the various MCS/P automated tools, did emerge during PW 96 to demonstrate the promise of digitization to greatly improve warfighter effectiveness. For example, in the Armor brigade, the engineer used the 3D terrain elevation tools, line of sight, and field of view to effectively optimize his units positions in relation to the OPFOR.

For key staff planning and coordination portions of the TDMP (Mission Analysis, COA Development and Analysis, and Wargaming) as well as some key staff tasks conducted during battle execution phases (Engineer operations, Tactical Fire Direction, and ADA), the MCS/P does not totally support "two-way" interactive parallel planning between higher and adjacent units as well as it might. The Critical software deficiencies in aiding collaborative staff performance included (1) the lack of an "event-driven Synch Matrix" tool, (2) the lack of automated OPORD and report generation tools (the Windows Desktop Display (WINDD) and networked file-server served this purpose), and (3) the lack of an effective map display (i.e., poor resolution, poorly read map terrain and awkward scaling tools) which resulted in the commanders using paper maps at the division level to do mission planning.. The commander's mental expectation from his or her past experience using a synchronization matrix was that it is an event driven process. Instead, the MCS/P required the commander to synchronize his plan according to an arbitrary time schedule. This lack of cognitive congruency resulted in the battlestaff using Windows Desktop Display software to manually develop their Synchronization Matrix. Over 65% of the MSF responses rated the synchronization matrix as being slower, not intuitive, and useful as paper-based, voice-communicated methods (Chi Square = 12.07, $p < .02$).

For collaborative staff planning, the MCS/P RCP overlays provided what many MSF staff members considered as simple Division level sketches, which very effectively flowed among the MSF. The RCP assisted the various MSF staff elements in sharing a "common picture" of the division battlespace. However, the RCP was only as timely (and, hence, accurate or relevant) as the "age" of the data sources used and the effectiveness of the TTPs followed for its development. Thus, the MSF staff generally considered the RCP to be of limited value because of the time lag inherent in its production. Finally, because of the MCS/P's small screen and its lack of resolution, collaborative planning within a cell was not easily accomplished because all parties could not view the MCS/P display with the needed detail. Therefore, much collaborative planning at various MSF staff elements still centered around the use of large paper maps. The majority of the MSF (60%) rated the MCS/P as offering only borderline support in the conduct of mission planning and analysis (Chi Square = 60.1, $p < .01$).

MCS/P capabilities proved very effective if time were constrained (such as in a time-constrained fragmentary order (FRAGO)). Automated graphic capabilities such as plotting enemy locations from databases, establishing unit Order of Battle, tracking high priority target artillery groupings, identifying key road networks, or factoring in visibility or elevation data became very effective tools for collaborative interaction between time stressed planning cells. On the other hand, because of the functional complexity of using MCS/P for quick time and space analysis in planning immediate actions to exploit windows of opportunity or eliminate unseen threats, some BDE staff elements found MCS/P less than optimum to execute these fast paced mission coordination and execution functions. Eighty two percent of the MSF that responded felt that the MCS/P did not support the staff or offered only borderline support in simultaneously analyzing courses of actions (COAs) (Chi Square = 15.9, $p < .01$). This was especially noticed during close battle. Instead, for exchange of key time-sensitive close fight information, voice, size, activity, location, unit, time, and equipment (SALUTE) reports were the preferred means of communication between higher and adjacent echelons. During the slower paced planning phases of the TDMP, the MCS/P "Post-it Notes" capability was considered an excellent tool for CCIR information exchange. However, the majority of the MSF staff were not well enough trained to routinely establish "selective filter alarms" for the MCS/P to automatically screen and display CCIR oriented messages over the changing phases of the battle. During the time sensitive collaborative monitoring of the close battle, many staff members considered the "Post-it Notes" process too slow and cumbersome for use. Additionally, the lack of effective user-set alarm selectivity caused some staff members to be inundated with messages so they simply disabled the MCS/P alarm function. In the case of the "Air Strike Warning", one ADA staff officer indicated that processing and overlaying the information on friendly units took so many key and mouse manipulations that the resulting information about unit vulnerability was generated too late to be useful for warning threatened units. Because of the complexity and inconsistency of the various functionality's resident in MCS/P, many close fight coordination efforts between various staff elements (e.g., TOC-A and Divarty) were done by voice because of the slow response time and process-intensive effort to get critical information from the MCS/P system. In summary, given these MCS/P shortcomings, the voice mode became the communication channel of choice for time-critical collaborative exchanges for many MSF staff members.

5 Summary

Based on the Performance Based Metrics, the MCS/P was somewhat effective in creating the MSF staffs' situational awareness and in portraying and communicating a timely and accurate relevant common picture (RCP). The MCS/P's performance in allowing the user to build and transfer databases was a strong point. However, there were some shortcomings to the system provided to the MSF. The wargaming tool did not work. The system was not stable, being prone to crashes for much of the SIMEXes due to lack of feedback. Editing and management tools for database records needed improvement, as well as several system usability and interface characteristics.

While individual performance of MCS/P-aided collaborative planning and execution tasks greatly improved from the initial attempts in SIMEX I, some shortcomings remained unchanged throughout the experimentation due mainly because of a lack of in-depth experience on MCS/P and MCS/P software deficiencies. This lack of experience resulted in the staff members spending large amounts of time trying to determine how to execute an MCS/P function rather than spending time performing critical staff functions with MCS/P. The MCS/P user's manual was ineffective and this greatly contributed to the training shortfall. The overall MSF staff skill levels on MCS/P software functionality continued to improve throughout PW 96, but some individuals never reached fully effective levels. Some individual users, with higher degrees of computer skills and expertise with the various MCS/P automated tools, did emerge during PW 96 to demonstrate that digitization has the potential to greatly improve the soldiers' warfighter effectiveness.

Because of the lack of both confidence in MCS/P and experience of the various functionalities resident in MCS/P, many close fight coordination efforts between various staff elements were done by voice because of the slow response time and process-intensive effort to get critical information from the MCS/P system. The voice mode became the communication channel of choice for time-critical collaborative exchanges for many MSF staff members.

References

Leedom, D., & Simon, R. (1995) *Improving Team Coordination*: A Case for Behavior-Based Training, Military Psychology, 7(2), pp. 109-123

Molich, R. & Nielsen, J. (March 1990) *Improving a human-computer dialogue: What designers know about traditional interface design. Communication of the ACM,* **33(3)**.

Nielsen, J. & Molich, R. (1990) Heuristic Evaluation of User Interfaces. *Computer Human Interface (CHI) 90 Proceedings.* pp. 249-255.

Prairie Warrior 96 Experimenting Agencies. (1996) *Prairie Warrior 96 Advanced Warfighting Experiment Final Report,* U.S. Army TRADOC Analysis Center (in press).

Smith, S. & Mosier, J. (August 1986) *Guidelines for Designing User Interface Software. Report MTR-10090.* The Mitre Corp., Bedford, MA.

INTENTIONALLY LEFT BLANK.

The Impact Of Field Of View On The Performance Of Some Infantry Tasks

Eugene Dutoit
D.Ayers, F.Heller, C.Holloway, K.McDonald, E.Redden
Dismounted Battlespace Battle Lab
Fort Benning, Georgia 31905

ABSTRACT

The Dismounted Battlespace Battle Lab (DBBL) conducted an experiment to determine the impact of field of view (FOV) of image intensification night sights on the capability of a unit and individuals to perform various Infantry related tasks. Three FOVs were investigated; 32, 40 and 60 degrees. The night vision sights were all monocular and mounted on the soldiers helmet. The experimental hypothesis / claim was that if the soldier is taught proper scanning techniques, the narrow field of view devices will provide the same operational capability as the larger FOV sights. The payoffs for using the smaller FOV goggles are; reduced weight carried on the helmet, increased image resolution and reduced hardware costs. Data were collected on a variety of Infantry tasks; for mounted and dismounted operations. The measures of effectiveness (MOE) were based on unit and individual performance. The methods of data analysis were primarily nonparametric, however parametric methods were also used and the "decisions" resulting from these two approaches were compared.

The purpose of this paper is to outline the analytic methods applied to experimental results obtained at Fort Benning, Georgia and compare the statistical decisions regarding specific measures of effectiveness (MOE) using nonparametric and parametric methods. The purpose of the experiment was to determine if there were any differences in soldier performance of some Infantry tasks when the field of view (FOV) of monocular night vision goggles (NVG) is varied (32, 40 and 60 degrees). This paper will focus on the analytical results obtained for each of the MOE and present the pertinent results as well as the statistical decision for each method of analysis.

The experimental hypothesis: if the soldier is taught proper scanning techniques, the narrow FOV devices will provide the same operational capability as the larger FOV devices. If this is true, then the payoffs to the Army will include; reduced weight on the soldier's helmet, increased image resolution and reduced hardware costs.

The system characteristics for the three NVGs are presented on the next page. Note that the weight of the 60 degree system is nearly twice that of the two other alternative systems.

Characteristic	32 and 40 Degree	60 Degree
Weight with batteries	.95 Lbs	2.1 Lbs
Focus range	25 mm to infinity	25 mm to infinity
On-axis resolution at optimum light level	1.3 CY/mr	1.2 CY/mr
Diopter focus	+2 to -6 diopters	+2 to -2 diopters
Exit pupil	10 mm @ 25 mm eye relief	12 mm @ 20 mm eye relief

SYSTEM CHARACTERISTICS

Overview of Training. In summary, the following steps were taken to train the test subjects. Each of the subjects was an Army soldier.

1. The test subjects were never told that the FOVs for the three goggle systems were different.
2. Each subject was taught how to focus each goggle and adjust the head harness.
3. Each subject walked the in-door Night Fighting Test Facility (NFTF) at Fort Benning. The facility has lanes established for the following environments; jungle, woodland, desert and urban. Each lane is approximately nine feet wide and 45 feet long.
4. Each subject also received additional training at the NFTF in these skills; boresighting each goggle, basic maintenance and firing an M16 rifle from the standing, foxhole and the prone firing positions.
5. Finally, each subject went to the out-door Buckner Range and was taught the preferred scanning techniques to use during the conduct of the experiment.

The following table provides the Infantry tasks and their related measures of effectiveness that were analyzed in this experiment.

Cross country dismounted movement	1. Number of navigation errors 2. Number of targets found 3. Navigation exercise time 4. Number of trips/stumbles
Cross country vehicle	1. Motorcycle exercise time 2. Ranger special ops vehicle; exercise time 3. Number of cones knocked down
Military operations in urban terrain (MOUT) performance	1. Time required to clear a room
Target engagement performance	1. Fraction of target detections 2. Fraction of targets hit

The experimental results and a summary of the statistical analysis for each of the ten measures of effectiveness listed in the table above will be addressed separately in this paper. The "statistical tools" used to analyze the experimental data are outline in the table below.

STATISTICAL TOOLS
* EXPLORATORY DATA ANALYSIS
* PARAMETRIC ANALYSIS OF VARIANCE
* NONPARAMETRIC (KRUSKAL WALLACE) ANOVA
* CHI-SQUARE (GOODNESS OF FIT AND CONTINGENCY)
* LOG-LINEAR ANALYSIS
* BINOMIAL PROBABILITY CALCULATION

Constraints and statements concerning the experiment and data analysis:

The subjects were initially assigned at random to each of the three NVG systems and the use of the goggles by the subjects was performed in a counter balanced procedure. However, the following list of caveats applies to this experiment and to the results.

a. There were no considerations of statistical power. This is consequence of the Advanced Warfighting Experiment (AWE) philosophy which is based on "looking for insights" as opposed to probabilistic decisions concerning experimental hypotheses.

b. In many cases data were obtained on tactical units instead of on individual soldiers. Unit analysis has an operational flavor and appeal that is hard to argue about. However, the unit analysis results in a "small sample" size. This in turn biases the experiment in favor of the null hypothesis (ie, no statistical differences).

c. An examination of the first chart of this paper clearly indicates that this experiment was a comparison between "systems" and FOV performance rather than a pure FOV comparison. The helmet mounted 60 degree FOV system was nearly twice as heavy as the other two helmet mounted systems and could have biased the results of the experiment.

d. Some of the MOE data elements were more subjective than desired. The measures of the "time to clear a room" and the count of the number of "trips and stumbles" were, in retrospect, rather subjective.

e. There was an unfortunate vehicle accident during the course of the experiment. As a result a soldier suffered a broken leg. This incident may have had some influence on the results of the remaining "vehicle" exercises.

RESULTS. The results of the data analysis will be presented for each MOE. It was not appropriate to use multivariate methods because many of the MOE had small numbers of observations. Nonparametric methods were used as the primary means of hypothesis testing. However, in several cases, the alternative parametric test was also conducted in order to determine if the statistical decision was conserved. The critical level of

significance was set at 10%. The results for each of the ten MOE are presented in the tables below.

Task: Cross Country Dismounted Movement.

MOE 1 is: Number of navigation errors.
Data obtained on a tactical team basis. Six teams or six observations per FOV.
Navigation error is defined as being off course by greater than five degrees.
Exploratory analysis indicated one outlier which was then removed.
FOV (degrees) 32 40 60 Average results 1.5 .83 1.00
Statistical Decision: Kruskal Wallace; P = .59 ANOVA; P = .51

Task: Cross Country Dismounted Movement.

MOE 2 is: Number of targets found.
Data obtained on a tactical team basis. Six teams or six observations per FOV.
MOE based on the number of targets found on the navigation course.
One team set of data were removed because the team was far off the course.
FOV(degrees) 32 40 60 Average results 1.25 2.00 1.25
Statistical Decision: Kruskal Wallace: P = .27 ANOVA: P = .26

Task: Cross Country Dismounted Movement.

MOE 3 is: Time to complete exercise(seconds)			
Data obtained on a tactical team basis. Six teams or six observations per FOV.			
Exercise time is defined as the time it takes the unit to walk the course.			
All observations included in the analysis.			
FOV (degrees)	32	40	60
Average results	85.2	68.7	79.8
Statistical Decision:	Kruskal Wallace; P = .39 ANOVA; P = .49		

Task: Cross Country Dismounted Movement.

MOE 4 is: Number of trips or stumbles.			
Data obtained on a tactical team basis. Six teams or six observations per FOV.			
The number of trips or stumbles were recorded for each unit.			
All observations were included in the analysis.			
FOV (degrees)	32	40	60
Average results	6.5	4.5	5.8
Statistical Decision:	Kruskal Wallace; P = .59 ANOVA; P = .88		

Task: Cross Country Vehicle; Motorcycle Exercise Time.

MOE 5 is: Motorcycle exercise time in minutes.			
Data obtained per motorcycle operator. Four to five operators per FOV.			
The time required to navigate the vehicle course was recorded.			
All observations were included in the analysis.			
FOV (degrees)	32	40	60
Average results	17.2	14.8	20.6
Statistical Decision:	Kruskal Wallace; P = .31 ANOVA; P = .20		

Task: Cross Country Vehicle; Ranger Special Operations Vehicle.

MOE 6 is: Ranger Operations Vehicle exercise time in minutes.
Data obtained per vehicle operator. Four to five operators per FOV.
The time required to navigate the vehicle course was recorded.
All observations were included in the analysis.

FOV (degrees)	32	40	60
Average results	20.9	19.7	22.3

Statistical Decision: Kruskal Wallace; P = .38
ANOVA; P = .69

Task: Cross Country Vehicle; Motorcycle and Ranger Vehicles.

MOE 7 is: Number of cones knocked over on the course by both motorcycles and the Ranger special operations vehicle.
Data obtained for each operator. Four to five operators FOV.
The data were collected/grouped and analyzed according to the following categories; FOV (three levels; 32, 40, 60 degrees), type of vehicle (two levels; motorcycle and Ranger) and side of vehicle which hit the cone (two levels, left and right).
All data included in the analysis.

These are the enumerated data:

	FOV			Type of Vehicle		Side of Hit	
	32	40	60	Moto	RSOV	Left	Right
	Number of Cones Knocked Over						
	34	29	19	12	70	33	49
P=		.32		.00		.27	

Hierarchical Log Linear Results are as follows:

* First order effects are adequate to explain the data. This is reflected in the methodology shown above for each of the categories.
* The "type of vehicle" is the driving factor. This is reflected above in the P value of .00 for type of vehicle. The ranger special operations vehicle had a significantly greater number of events.
* FOV is the weakest factor. This reflected in the P value of .32 presented above.

Task: Military Operations in Urban Terrain.

MOE 8 is: Time to clear a room (minutes).
Data obtained on a tactical team basis. Six teams or six observations per FOV.
MOE is determined from the time the team first enters a room until all the enemy is determined to be killed. This is a subjective determination by the subject matter experts and is considered to be "weak".
Exploratory analysis indicated one outlier which was removed.

FOV (degrees)	32	40	60
Average results	1.26	1.03	1.15

Statistical Decision: Kruskal Wallace; P = .83
ANOVA; P = .63

Task: Target Engagement Performance.

MOE 9 is: Fraction of available targets detected.
Data were obtained for each individual soldier. Twenty soldiers were used in the experiment. The use of each FOV by each soldier was randomized. It was assumed that each shot at a "target" equaled a detection. There were twenty target opportunities per soldier per FOV; or a total of 400 target opportunities per FOV. Information on false detections was not available.
The fraction of available targets detected for each FOV = Number of detections (or shots) divided by 400.
All data were included in the analysis.
The results for each FOV are: Prob Detection for FOV of 32 degrees = .558. Prob Detection for FOV of 40 degrees = .558 (this is not a typo). Prob Detection for FOV of 60 degrees = .543.
Statistical Decision: There was no statistically significant difference in performance between the three FOVs. There was overlap between the three 95 % confidence intervals computed about each of the point estimates cited above.

Task: Target Engagement Performance.

MOE 10 is: Fraction of detected targets that were hit. This was a different range than discussed above for MOE 9.
Data were obtained for each individual soldier. Twenty soldiers were used in the experiment. The use of each FOV system by each soldier was randomized. Each soldier was presented forty targets per FOV system in this target rich environment. Therefore each FOV system was exposed to,(20 X 40), 800 target opportunities for detection and engagement. The targets appeared to "pop-up" at random but they were actually programmed to appear random.
The fraction of target hits for each FOV = Number of hits divided by the Number of detections.
All data were included in the analysis.
The results for each FOV are:

	Estimate	95% Confidence Interval
Prob Hit for 32 degree FOV	= .487	.452 -- .522
Prob Hit for 40 degree FOV	= .350	.318 -- .382
Prob Hit for 60 degree FOV	= .367	.333 -- .400

Statistical Decision: The probability of hit for the 32 degree FOV system is greater than the other two systems.

SUMMARY. The results for each of the ten MOE were presented in tables above and should be able to speak for themselves. The following list of results is intended to be a summary of the results and conclusions across all of the MOE. Some of these general statements have already been discussed.

1. The experimental hypothesis of equal effectiveness using the different FOV systems for the selected Infantry tasks is supported. There was no statistically significant difference in performance for nine out of ten tasks. In MOE ten, the probability of hit for the 32 degree FOV system was statistically better than for the other two systems. However, it needs to be repeated that the small samples (on a unit or tactical team basis) will result in low statistical power. It was also evident that there were physical differences between the systems (such as weight) and pure FOV was confounded with other system parameters.

2. Although there was only one case were the differences in FOV performance were statistically significant (MOE 10), a careful examination of the tables for each MOE shows that the 40 degree FOV system is the "best, or tied for best", eight times out of

ten (Reference MOE 1,2,3,4,5,6,8 and 9). The probability that any single one of these FOV systems would be "best, or tied for best", eight out of ten times under the null hypothesis of no difference in performance is .003. This is a rather interesting result. The 40 degree FOV system is the same weight as the 30 degree FOV system and should be more comfortable to wear on the helmet than the larger 60 degree FOV system.

3. The 60 degree FOV system was "best" for MOE 7; fewer cones were knocked down on the driving course using this system.

4. The Ranger special operations vehicle was involved in a significantly greater number of events (reference MOE 70.

5. The statistical decisions were consistent when both nonparametric and parametric methods were applied to the data. This result is not surprising when the small sample sizes are considered.

INTENTIONALLY LEFT BLANK.

AUTOMATED EMPIRICAL EVALUATION OF THE FACT EXCHANGE PROTOCOL

Maria C. Lopez, Ann E. M. Brodeen,
George W. Hartwig, Jr. and Michael J. Markowski
U.S. Army Research Laboratory
Aberdeen Proving Ground, Maryland 21005–5067

ABSTRACT

Decentralized battlefield command and control requires reliable and timely distribution of information. At present, distribution of digital information is limited by the low–bandwidth noisy channels inherent to combat net radios and heavy traffic demands, forcing commanders to make decisions from less than timely information. In the ideal communications network, each node would be smart enough to monitor network performance and, when necessary, adapt itself to make better use of the available bandwidth. The adaptive network node would employ a decision algorithm to modify configuration, routing and protocol parameters based on measured network performance statistics and system requirements. Our research addresses the effects of noise and interference on communications channels and construction of network protocols that will be effective on the modern battlefield. The approach emphasizes use of actual hardware and controlled experimentation to explore alternative protocols. This paper describes a controlled laboratory experiment in which messages were passed over a communications network using the combination of the Fact Exchange Protocol (FEP), the Tactical Data Buffers (TDBs) and Single Channel Ground and Airborne Radio System (SINCGARS) Combat Net Radios (CNRs). It also describes the suite of software to automatically execute the test design, and collect and apply preliminary data reduction procedures to baseline performance data for the prototype communications network.

BACKGROUND

The primary means of communications at low–echelon fighting units has been and continues to be voice data transmitted by CNRs. Gradually, a requirement for digital data transmission is being inserted into the mission profile. Digital transmissions allow for compression and forward error correction and provide the ubiquitous computer with the information it requires. With this increasing requirement for digital transmissions, problems arise.

Modern combat net radios are typically line–of–sight, Frequency Modulation (FM), low power instruments designed specifically for use at short range. Their bandwidth is very limited, typically 1200–2400 bits per second (bps), although recent improvements in modem technology have pushed these numbers as high as 16 kilobits per second (kbps). These radios are commonly assembled into a single hop network of 6 to 12 users. Their effective use to date is testimony to the redundancy of the human language and the ability of the human brain to extract meaningful data from a noisy signal.

Our research addresses the effects of noise and interference on communications channels and construction of network protocols and procedures that will minimize delay and maximize throughput on the modern battlefield. The networks that are of particular interest to us have nodes with high computing power but weak, noisy, shared communications links. For this reason, our approach to communications emphasizes intelligent processing at each node to limit the amount of information that must be passed along the communications channel. Each node is assumed to act independently to improve the effectiveness of the information exchange between nodes. Such a system of controls requires that each node be able to monitor the network traffic; decide whether performance is inadequate; and, if so, make an appropriate adjustment to the protocol.

A series of controlled experiments is being conducted to determine which communications protocol parameters and structural assumptions have the greatest impact on selected performance measures. To accomplish such an objective, it is required that a group of computers serving as battlefield nodes be synchronized, network parameters be initialized prior to each run, and collected data be made conveniently accessible to the user. As a result, software that performs the necessary tasks with minimal user intervention was developed.

TEST CONFIGURATION

There are three nodes, each of which is a SPARCbook 3.[1] Each contains a communications protocol and a scenario driver. The communications protocol includes data collection functions to log the sending and receipt of messages

Approved for public release; distribution is unlimited.

and acknowledgements (ACKs) as well as information on the queues. The scenario driver provides the communications loading. The nodes are connected, via ethernet, to a SPARCstation 20[2] that serves as the data storage and control node. The nodes are connected to SINCGARS CNRs via TDBs, a modem between the radios and the terminal equipment. Resistor loads are used as antennas to reduce the transmission range.

The TDB interfaces with the computer using RS–232C, and with the SINCGARS using MIL–STD–188(C). Two processing steps are performed to input data to the TDB: 1) any formatting bits, such as start, stop, and parity, are removed so that transmission time is not expended by unnecessary data; 2) the data are stored until the TDB can access the network. The storage capacity is 24 kilobytes. Storing the input data avoids collisions between incoming and outgoing data.

The TDB may process the data to be sent in a number of ways depending upon the setting of various internal and front panel switches. In the simplest mode nothing is done to the data and it is output at the raw data rate of the TDB of 16 kbps. The simplest processing that can be selected uses the Bose–Chandhuri–Hacquenghem (BCH) protocol for error detection/correction. Characters are coded in 4 byte groups at a 48/32 rate. In other words, each 32 bit or 4 character block becomes 48 bits after encoding. This encoding reduces the effective throughput to 10.66 kbps. Finally, three modes of forward error correction may be requested. This error correction algorithm consists of retransmitting multiple copies of the data. The first setting causes no forward error correction to be done, i.e., the data is sent once and the effective throughput is still 10.66 kbps. The next setting causes the data to be repeated 5 times and interleaved in a manner designed to spread out burst errors. The effective throughput at this level of redundancy is 2.133 kbps. The last setting causes the data to be repeated 13 times resulting in an effective throughput of 820 bps. Forward error correction with a redundancy of 5 was selected for this experiment.

The receiving TDB performs the appropriate level of de–interleaving. In those cases where data is repeated, it uses majority voting to resolve differences between redundant blocks, and does BCH decoding resulting in a block of 4 characters. If, for any reason, the characters cannot be identified, the damaged 4 character block is replaced in the output stream with the four characters "@@@@". The data are then passed to the storage buffer where formatting bits are reinserted and then output on the RS–232C line to the data processing device. For more details refer to Harris.[3]

Figure 1 illustrates the test configuration.

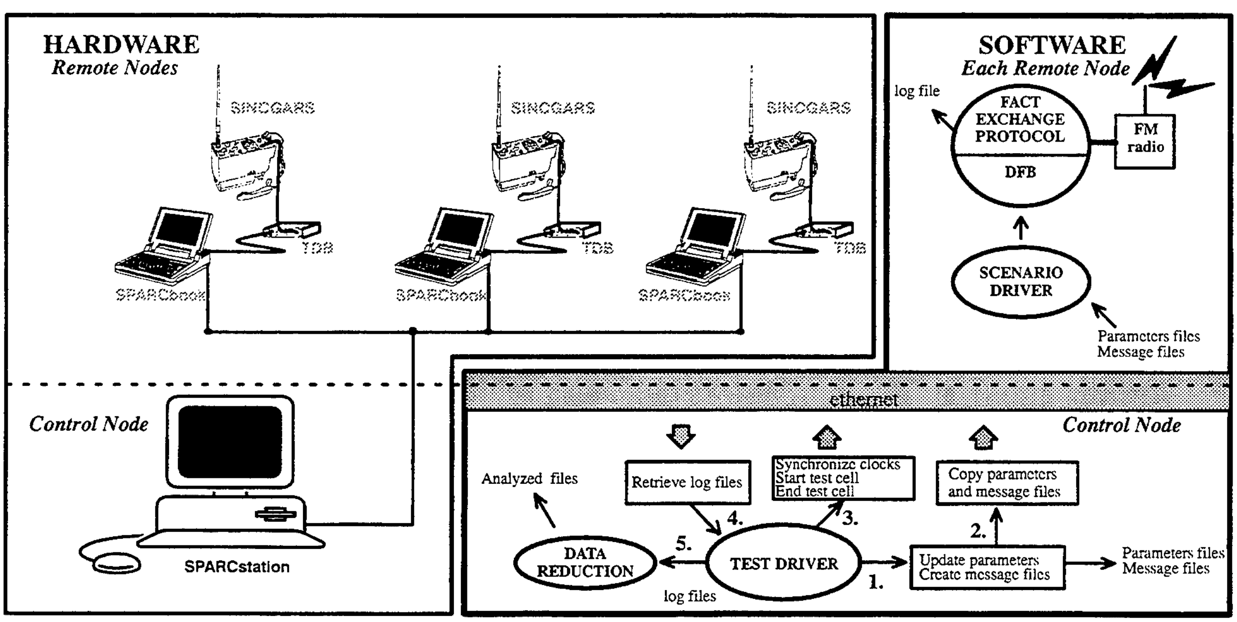


Figure 1. Test Configuration

SOFTWARE CONFIGURATION

The software consists of four parts: the test driver, the data reduction software, the scenario driver, and the communications software.

The **test driver** is a menu–driven user interface written in the C programming language,[4] uses X Windows[5] and Motif,[6] and runs under a UNIX[7] operating system. It coordinates all tasks necessary to execute the experimental design. Prior to the test driver existence, the experimental design for similar tests was executed manually, requiring extra time for setup and the possibility of errors during the initialization phase of a test cell.

Among its tasks, the test driver generates messages for the scenario driver, updates the factor–level combinations, distributes the information to the nodes, and synchronizes the nodes' clocks. In addition, it starts and ends each test cell, retrieves all log files from the remote nodes for storage on the control node, and computes network statistics. To minimize input errors, the test driver runs all experimental combinations without human intervention. The software is capable of executing independent replications of the design matrix automatically, with each replication using different random numbers, starting in the same initial state, and all statistical counters reset to zero.

The test driver reads information contained in text files to initialize values that may vary depending on the experimental design. These text files contain values that need initialization prior to the test cell such as: factors and levels of interest; the number of replicates for each test cell; the number of replicates for the center point; the random number seeds to generate the desired message sets or scenarios; the number of tries for each message; node identification string; and the length of each run. Other values that are initialized are the names of the directories into which the software will store the data, the directories where executable binary files are located, and values that are used by the data reduction software. The text files used for initialization may be modified either by editing the files prior to running the test driver or by menu selection before executing the experimental design.

The communications and scenario driver software on the remote nodes have their own input files; these also need to be updated prior to each test cell. The control node has a copy of these input files, referred to as template files, which the test driver updates and copies onto the remote nodes. Template files are used whenever part of a file needs to be modified more than one time during the test run. Examples of this kind of file are the capabilities input file (cif_nodename) loaded by the communications software to initialize the nodes' id, the window size and retry time–out (Figure 2a), and the nodename## file from which the scenario driver gets the message information to load messages into the communications software.

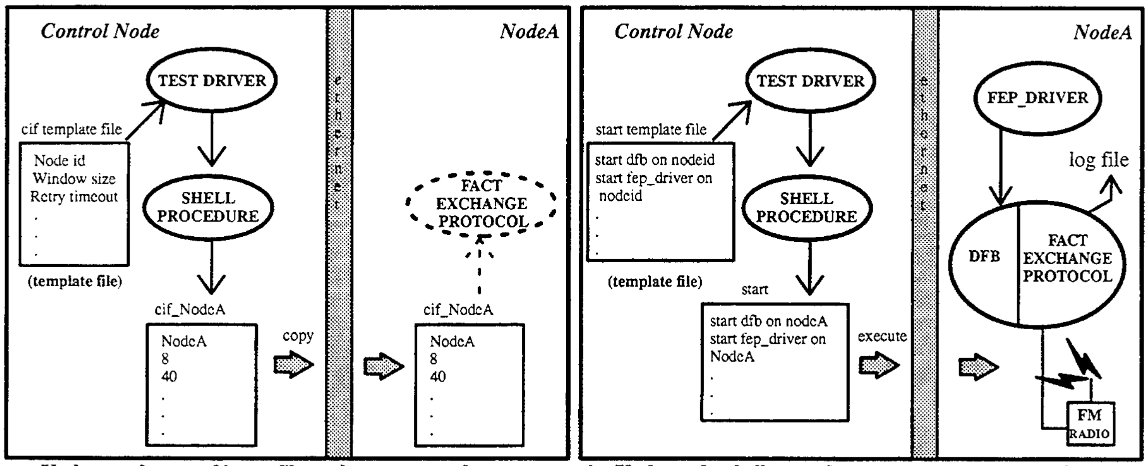
The test driver invokes UNIX shell procedures to execute tasks on the remote node such as synchronizing clocks, starting and ending the execution of a test cell (Figure 2b), as well as on the control node, such as copying files to the remote nodes (Figure 2a) and retrieving log files from the remote nodes.

During the execution of a test cell, each node collects data in a log file local to that node. The log files contain time tagged information on the messages and ACKs sent and received, as well as information on queues. The **data reduction** software is a set of C programs that reformats log files and computes network statistics. The test driver executes UNIX shell procedures to invoke the data reduction software. The shell procedures that contain node information are updated using template files. The output of the data reduction software is formatted in a fashion suitable for statistical analysis.

The **scenario driver** is a C language application that reads a file of time tagged, preformatted message strings and forwards them to the DFB at the appropriate times.

The **communications software** is a C language application composed of a freeform database management system called the Distributed FactBase (DFB), which communicates with the other DFBs via the FEP. An important concept implemented in the DFB is the ability to automatically initiate predefined actions (rules) upon receipt of new information. These rules ensure that only significant data (as defined by the commander and staff) are transmitted.[8] The FEP is a tactical transport layer protocol that communicates information quickly, concisely, and reliably over unreliable, low–bandwidth CNRs. It is designed to be a connectionless, reliable protocol (guarantees delivery of messages within certain parameter limits) that utilizes multicast, overhearing, and other techniques to minimize radio transmissions.[9] A data collection function is provided by the DFB to log information on messages, including ACKs, transmitted and received.

Figure 3 illustrates the software configuration.



a. Update and copy of input file to the remote node.

b. Update of a shell procedure to start a process on the remote node.

Figure 2. Template File with Shell Procedure Interaction

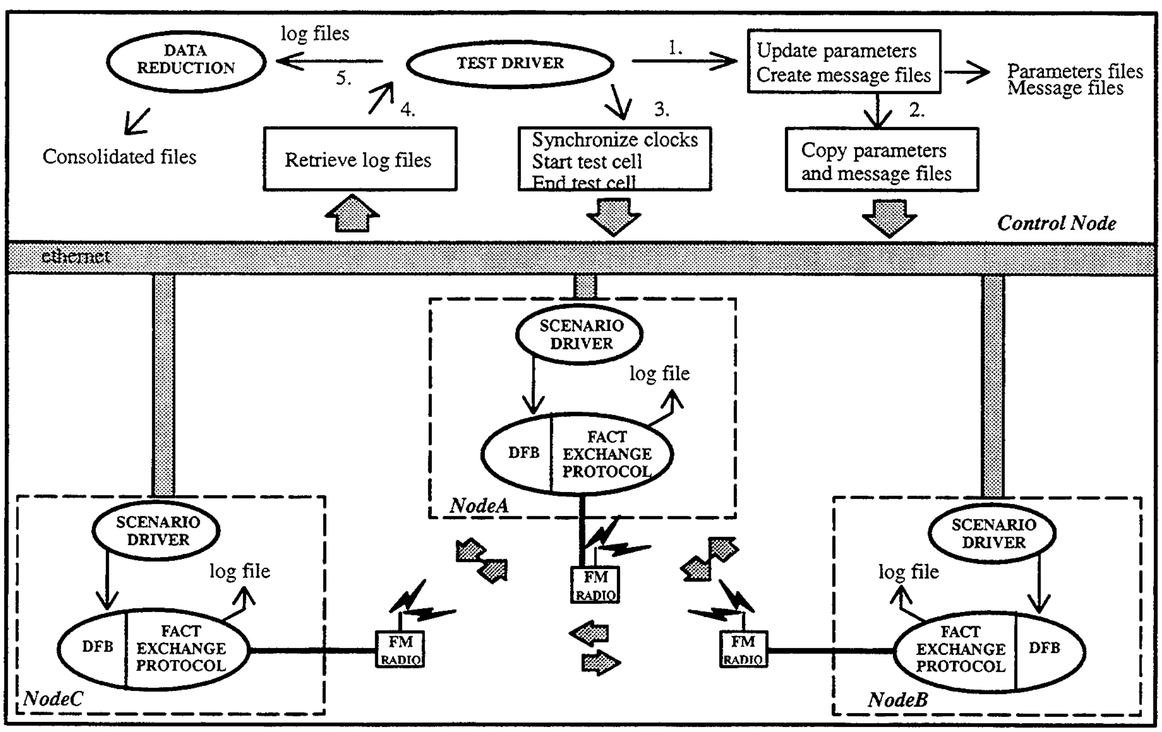


Figure 3. Software Configuration

EXPERIMENTAL DESIGN AND ANALYSIS

EXPERIMENTAL DESIGN

Experimental design provides a means of deciding before any runs are made which particular configurations to examine so that the desired information can be collected with the least amount of testing. Carefully designed experiments are much more efficient than a "hit–or–miss" sequence of runs in which a number of alternative configurations are unsystematically tried just to see what happens.

When the number of factors is moderate, a factor–screening strategy, such as a factorial design, might be able to indicate which factors appear to be important, and more to the point, which factors are irrelevant and can be simply fixed at some reasonable level and omitted from further consideration. The software developed in–house currently supports the fully automated execution of a modified 2^k factorial design. The four factors selected for testing, retry time-out interval, window size, message arrival rate, and message length, are ones that can be easily modified.

Two levels of each factor were tested with each of 2 levels of every other factor yielding 16 test combinations. The levels of each factor are listed below:

1. Retry time–out (time in seconds a host waits for an ACK before retransmiting the message)
   10
   40
2. Window size (number of messages allowed to be sent per host without waiting for an ACK)
   8
   50
3. Message arrival rate (per one hour test cell)
   200 per node
   600 per node
4. Message length (in characters)
   80
   240

Past experimentation with actual hardware and a tactical communications protocol illustrated that network behavior is nonlinear in nature.[10] A potential concern with the use of two–level factorial designs is the assumption of linearity in the factor effects. That is not to say that a 2^k system requires perfect linearity – this system works quite well even when the linearity assumption holds only very approximately. However, to provide protection against anticipated curvature in the response data, the 2^k design was augmented with five center points (corresponding to a retry time–out of 25 seconds, a window size of 29, an arrival rate of 400 messages per node, and a message length of 160 characters). The entire experimental design was replicated three times.

BIRNBAUM–HALL TEST FOR DIFFERENCES AMONG NODES' TIME TO SUCCESS

We wish to determine whether the distribution functions for the time to success data for the three experimental nodes are identical, especially in light of the fact that the hardware representing one of the nodes was equipped with greater memory. The Birnbaum–Hall test has been selected for several reasons: the data consist of exactly three independent samples, each of size $n = 63$; the random variable, time to success, is continuous making this an exact test; and, most importantly, the test is consistent against all alternatives.[11]

The null hypothesis is that there is no difference in the probability distributions of time to success among the three nodes, and the alternative is that a difference exists between at least two of the distributions. Although not shown here, the greatest vertical distance between any two of the empirical distribution functions occurs at a time to success of 304.1 seconds. This distance is $3/63 = .0476$. The critical region of size $\alpha = .05$ corresponds to all values of the test statistic greater than .2948, the large sample approximation for the .95 quantile from tables for the Birnbaum–Hall

statistic for n > 40. Therefore, there is not sufficient evidence to reject the null hypothesis, and we conclude the nodes do not differ with regard to the probability distributions of time to success.

Given the nodes appear to exhibit similar response behavior, and that performance of an individual node is not of singular interest, the data for the individual nodes will be combined into a single collective set for further exploratory analysis.

EXPLORATORY DATA ANALYSIS

Graphics is both a powerful exploratory data analysis tool for obtaining insight into the structure of data and a diagnostic tool for confirming assumptions or, when assumptions are not met, for suggesting corrective actions.

Many important properties of the distribution of a data set are conveyed by the quantile plot, including the median, quartiles, interquartile range, and other quantiles of interest, as well as information about the local density of the data and symmetry.

A preliminary look at the aggregate set of the nodes' time to success data is provided by the quantile plot in Figure 4. For this empirical data set, we see that the median is about 70 seconds and that a large fraction of the observed values lies between 25 seconds and 100 seconds. The longest time to success is in the neighborhood of 900 seconds, with a total of 36 observations greater than 200 seconds.

The data exhibit remarkably flat behavior below the .82 quantile, indicative of the local density, or concentration, of the data. This is revealed on the quantile plot by the string of nearly horizontal points.

The quantile plot may also be used to examine the data set for symmetry. If the data were symmetric the values in the upper portion of the plot would stretch out toward the upper right quadrant in the same fashion as the values

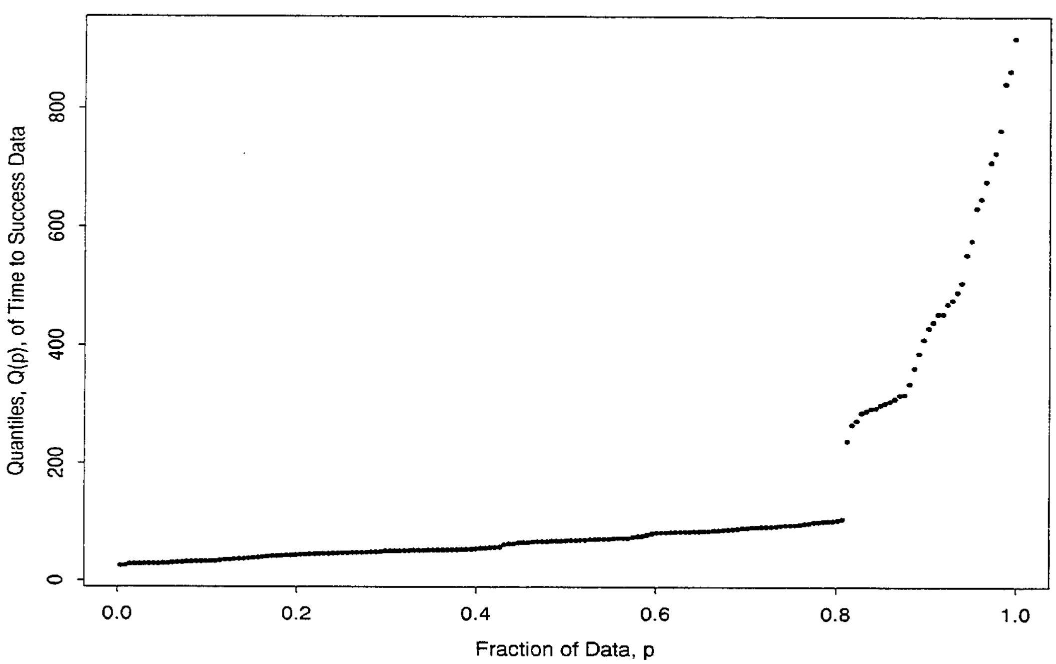


Figure 4. Quantile plot of the time to success data.

in the lower half stretch out toward the lower left quadrant. The observations in Figure 4 are skewed toward large values. Small values are tightly packed together; the large values stretch out and cover a much wider range of the measurement scale. The skewing increases dramatically as we go from small to large values, resulting in a strongly convex pattern. This is anticipated with network delay data.

Figure 5 displays the frequency of messages acknowledged as a function of try number for all 63 test cells. The communications protocol dictates that once a message is sent, if it is not acknowledged it is retransmitted. Each transmission was considered a "try". In this experiment, the protocol was configured to retransmit up to two times, yielding a total of three possible tries to transmit one message. The message was discarded if an ACK was not received after three tries.

From Figure 5, one can see that more than 50% of the messages either failed, i.e., not acknowledged within 3 tries, or were never transmitted due to the window size being full, causing the messages to literally be trashed. The trend exhibited by this distribution of messages is a mirror image of what should be generated by a network process under control. The information extracted from this plot was enough to warrant further investigation of the FEP and the DFB and halt further testing and analysis.

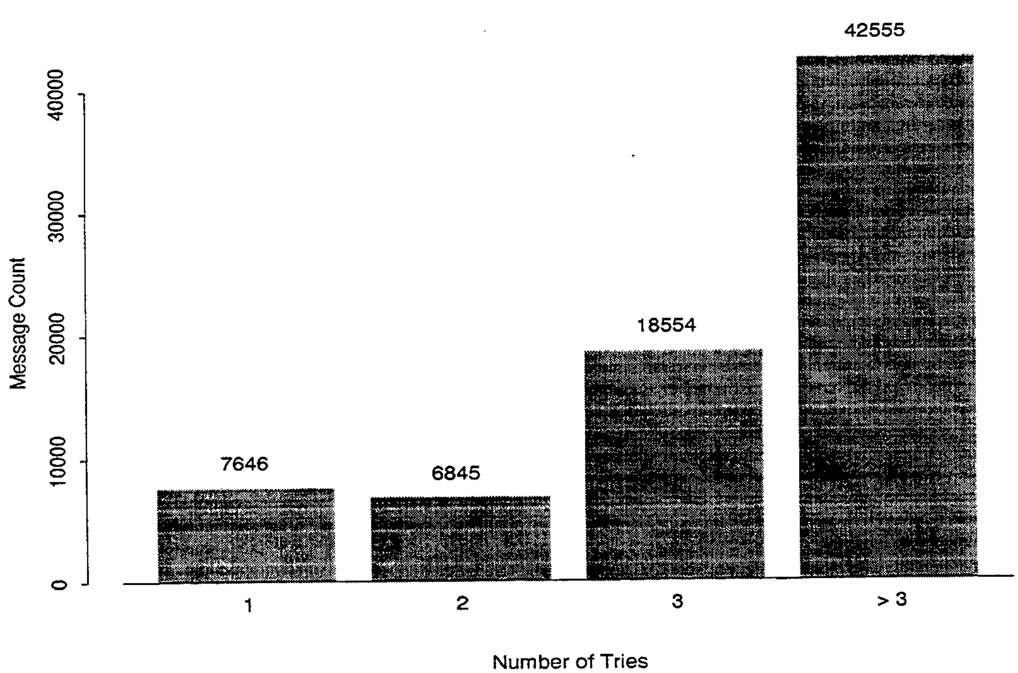


Figure 5. Distribution of messages by try number.

SUMMARY

The major problem identified by the pilot test was the FEP's failure to match outstanding messages with returning ACKs. This problem arose only when several messages were awaiting ACKs and resulted in the number of outstanding messages growing until eventually the window size was exceeded. This failure to match ACKs had two effects: 1) each message was transmitted the maximum number of retries greatly reducing total throughput; 2) once the window size was exceeded all transmissions were stopped.

The template files are useful in simplifying the programmer's job when the experimental configuration requires modification. Their use allows fast and easy modification to the experimental configuration since the input is not "hard wired" into the code. For instance, if the number of nodes needs to be increased or decreased, the programmer modifies the input text files containing node information and the updates on the remote software take place during the test driver initialization phase.

Because the test driver is of a general nature, it can be used in a variety of situations to run experiments in a distributed UNIX environment.

It is anticipated that future experiments can be automated to consider more complex communications protocol modifications. Automating the process reduces the chance of operator error and simplifies the execution of the experimental design.

REFERENCES

1. SPARCbook 3 Series Technical Reference Manual. Austin, TX: Tadpole Technology Inc., 1994.
2. SPARCstation 20 System Specifications Manual. Mountain View, CA: Sun Microsystems Inc., 1994.
3. Harris RF Communications RF–3490 Digital Data Buffer Instruction Manual. Rochester, NY: Harris Corporation, 1990.
4. Kernighan, B. W., and D. M. Ritchie. The C Programming Language. 2nd Edition, Englewood Cliffs, NJ: Prentice–Hall Inc., 1988.
5. Nye, A. Xlib Reference Manual. Volume 2, 3rd Edition, Sebastopol, CA: O'Reilly & Associates Inc., 1992.
6. Ferguson, P. M. Motif Reference Manual. Volume 6B, 1st Edition, Sebastopol, CA: O'Reilly & Associates, Inc., 1994.
7. McGilton, H. and R. Morgan. Introducing the UNIX System. New York, NY: McGraw–Hill Book Company, 1983.
8. Chamberlain, S. C. "The Information Distribution System: IDS – An Overview." BRL–TR–3114, Ballistic Research Laboratory, Aberdeen Proving Ground, MD, 1990.
9. Kaste, V. A. "The Information Distribution System: The Fact Exchange Protocol, A Tactical Communications Protocol." BRL–MR–3856, Ballistic Research Laboratory, Aberdeen Proving Ground, MD, 1990.
10. Kaste, V. A., A. E. Brodeen and B. D. Broome. "An Experiment to Examine Protocol Performance Over Combat Net Radios." BRL–MR–3978, Ballistic Research Laboratory, Aberdeen Proving Ground, MD, 1992.
11. Conover, W. J. Practical Nonparametric Statistics. 2nd Edition, New York: John Wiley & Sons, Inc., 1980

ANALYSIS OF SYNTHETHIC PROPORTIONS

Carl T. Russell
TEXCOM Experimentation Center
Fort Hunter Liggett, California 93928-8000

ABSTRACT

Continuous data lend themselves easily to graphical display, but analogous displays for discrete data such as hit/miss data are not so readily available. Nominal logistic regression can produce an estimate of success from the underlying regression model for each cell in the underlying contingency table. Since logistic regression uses maximum likelihood to fit logarithms of odds ratios, the estimates produced are strictly between zero and one even for cells with only one observation. Thus the underlying model enables one to transform discrete data into more nearly continuous "synthetic proportions" for analysis. Ordinary least squares regression can then be used to manipulate estimates into useful marginal estimates. Alternatively, graphical methods such as dotplots or boxplots can usefully display distributions of the synthetic proportions. Examples from small arms hit/miss data are used to illustrate promising techniques.

INTRODUCTION

This paper grew out of work done early in 1996 at the U.S. Army Test and Experimentation Command (TEXCOM), Experimentation Center (TEC), Fort Hunter Liggett, California. A proposed new sighting device for the M16 rifle and M4 carbine was tested. This device was supposed to improve the speed at which soldiers could fire on targets without degrading the probability of hitting the targets. The experimental design used was a classical sort of design but with lots of nesting. The sighting device had slightly different versions for the M16 and the M4 WEAPONs, and the standard "iron" sight was included as a baseline. Two different manufactures submitted candidates, giving a total of three SIGHTs (CANDA, BASELINE, and CANDB), considered to be nested in WEAPON. Twenty soldiers (ROSTER) executed six firing tables (TABLE, labels TAB1-TAB6 but actually corresponding to NBC, wide view, standard record fire, etc.) which consisted of firing rounds at various targets and range bands (RANGE, bands from 50 m to 300 m) which varied by TABLE (therefore nesting RANGE in TABLE). Soldiers fired a total of 18,960 shots (including some multiple shots at the same targets) under 3890 combinations of TABLE, RANGE, ROSTER, WEAPON, and SIGHT. Between one and eleven shots were fired under each combination of conditions, and both times of shots (from audio) and number of hits were recorded for each combination of conditions. Analysis of the time data was relatively easy using ordinary Analysis of Variance (ANOVA), and in the end the analysis could be easily displayed using boxplots without even referring to the ANOVA (see Figure 1). That analysis will not be discussed further in this paper. Instead, this paper discusses analyses of the hit/miss data which also yield graphical presentations.

ANALYSIS

Figure 2 shows a simple attempt to produce boxplots of hit/miss data. Even though the horizontal plot position of each data point is "fuzzed" by adding random error to alleviate overplotting, the plot is unhelpful with this much data. Thus a more sophisticated approach is needed.

Nominal logistic regression is an analog to ANOVA for hit/miss data. The "odds ratio" (ODDR) corresponding to a test condition is the ratio of the probability of hit to the probability of miss under that condition; that is,

$$\text{ODDR} = \text{Prob}[Y=\text{Hit}]/\text{Prob}[Y=\text{Miss}]. \tag{1}$$

In this paper, ODDR is also used empirically and somewhat ambiguously to refer to the the ratio of the proportion of hits to the proportion of misses under a particular condition. The logarithm of ODDR, ln(ODDR) has the nice symmetric and asymptotic properties desirable for classical linear model building:

$$\begin{aligned} &\ln(1/\text{ODDR}) = -\ln(\text{ODDR}) \\ &\ln(\text{ODDR}) \rightarrow -\infty \text{ as Prob}[Y=\text{Hit}] \rightarrow 0 \\ &\ln(\text{ODDR}) \rightarrow \infty \text{ as Prob}[Y=\text{Hit}] \rightarrow 1. \end{aligned} \tag{2}$$

Approved for public release, distribution is unlimited.

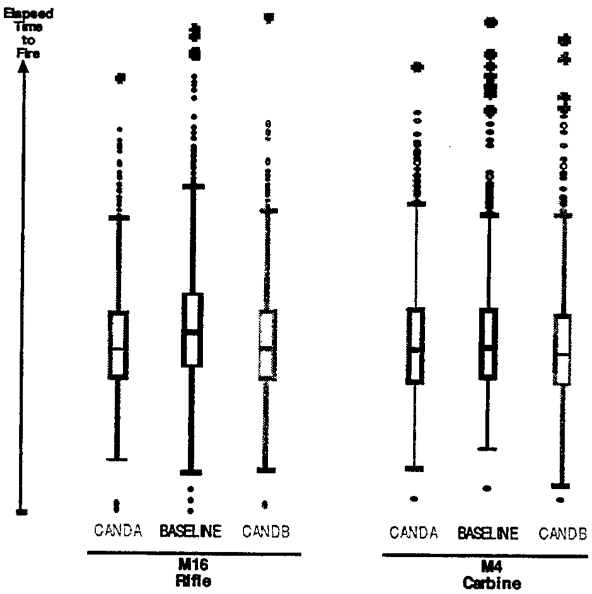


Figure 1. Boxplots of firing times.

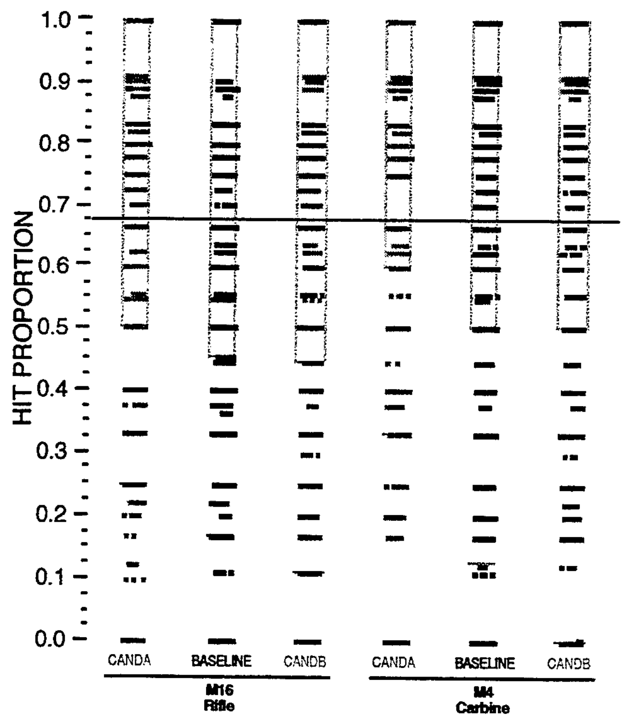


Figure 2. Boxplots of hit proportions (not very helpful).

Nominal logistic regression iteratively fits a multiplicative model for changes in ODDR through the loglinear model $\ln(ODDR) = X\beta$. Two simple exponential formulas then let one get back to estimates of hit and miss probabilities from the estimates of ln(ODDR):

$$\mathrm{Prob}[Y{=}\mathrm{Miss}] = 1/(1 + e^{\ln(ODDR)})$$

$$\mathrm{Prob}[Y{=}\mathrm{Hit}] = e^{\ln(ODDR)}/(1 + e^{\ln(ODDR)}). \qquad (3)$$

In this paper, nominal logistic regression is used to analyze the hit/miss data using the model reflected in Table 1. Because of the relatively complicated nesting, the factor of interest (SIGHT[WEAPON]) could not even start to be addressed until the other more influential factors of TABLE, RANGE and ROSTER and their interactions (or non-interactions) with WEAPON had been accounted for. Not surprisingly with such a large amount of data, some "statistically significant" effects involving SIGHT turn up, but they are clearly small compared to those of the more

Table 1. Statistical Summary of Hit Performance

Source	DF	ChiSq	Prob>ChiSq	ChiSq per DF
TABLE	5	187.73	0.0000	37.5
RANGE[TABLE]	26	2269.29	0.0000	87.3
ROSTER	19	550.70	0.0000	29.0
WEAPON	1	0.04	0.8399	0.0
TABLE*WEAPON	5	4.63	0.4631	0.9
WEAPON*RANGE[TABLE]	26	33.94	0.1366	1.3
WEAPON*ROSTER	19	65.00	0.0000	3.4
SIGHT[WEAPON]	4	3.76	0.4401	0.9
TABLE*SIGHT[WEAPON]	20	46.03	0.0008	2.3
SIGHT*RANGE[TABLE,WEAPON]	104	148.53	0.0027	1.4

Multiway contingency table analysis was performed on hit/miss data (18,960 shots) using nominal logistic regression as implemented in the SAS® JMP® statistical package (version 3.1). The overall model had a ChiSquare of 5079 with 229 DF, which is highly statistically significant.

influential factors. Nevertheless, some of the apparent effects could be operationally important. If the data were continuous a likely next step would be to look at the Least Squares Means (LSMs) because those are what are really being tested in Table 1. Fortunately, an analog to LSMs for count data can easily be obtained from computer statistics packages such as the SAS® JMP®[1] package which was used for this analysis.

AN ANALOG TO LEAST SQUARES MEANS

Once the fit ln(ODDR)=Xβ is obtained from logistic regression, one can theoretically string the vector β of model parameter estimates together just like SAS PROC GLM or JMP does in OLS regression to obtain LSM's. If you've ever tried to do that, you've undoubtedly found that a computer does a lot better job than a person. Luckily, JMP offers to produce and retain estimates of ln(ODDR) for each of the 3890 rows in the underlying contingency table which in this case produces the estimates portrayed in the giant linear combination in Figure 3. Call this

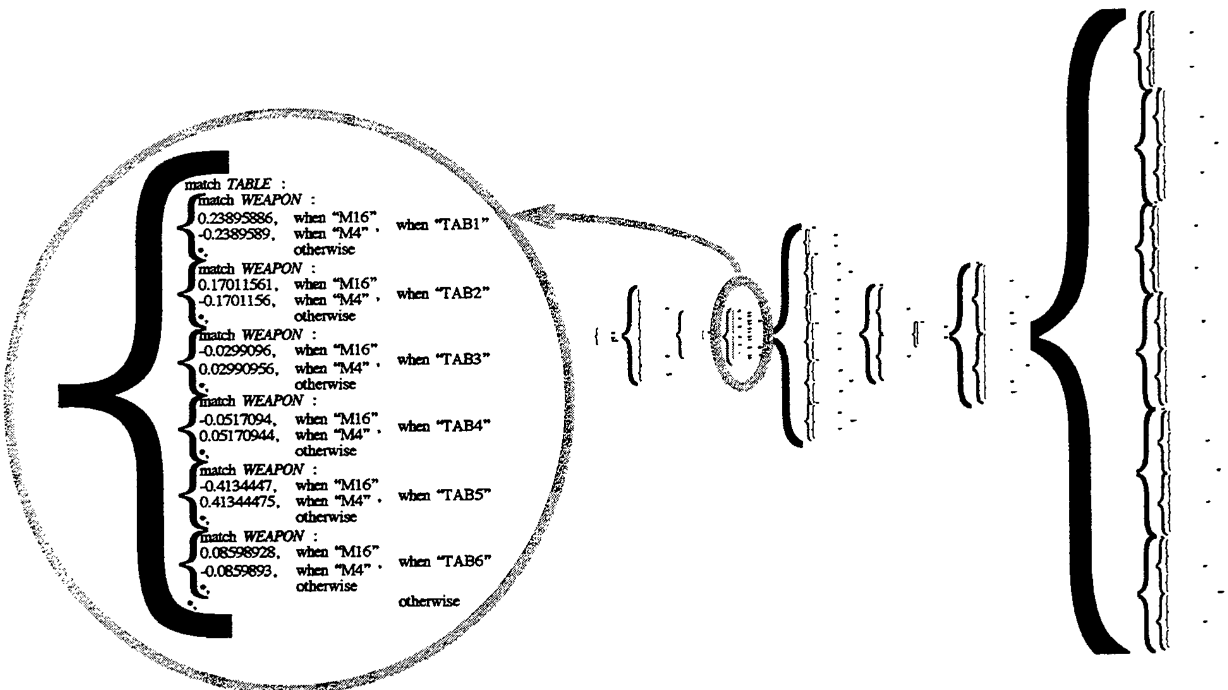


Figure 3. Example of ln(ODDR) estimates for each row of the underlying contingency table.

vector of estimates λ, and consider what happens when OLS regression is used to fit λ using the same **X** model used to obtain λ. Except for computational error, the fit will be perfect since OLS regression is simply undoing the

Table 2. LSMs from OLS regression on ln(ODDR) using the same model as in logistic regression (extract).

Level	LSM ln(ODDR)	Std Error
...	...	...
[TAB1,M16]075,CANDA	3.02	0
[TAB1,M16]075,BASELINE	3.44	0
[TAB1,M16]075,CANDB	9.26	0
[TAB1,M16]200,CANDA	1.54	0
[TAB1,M16]200,BASELINE	0.96	0
[TAB1,M16]200,CANDB	0.84	0
[TAB1,M16]300,CANDA	0.53	0
[TAB1,M16]300,BASELINE	0.14	0
[TAB1,M16]300,CANDB	-0.42	0
[TAB1,M4]075,CANDA	3.48	0
...	...	...

perfect linear fit coded into λ. So all statements concerning statistical significance are meaningless, but the estimates LSMs are fine, and they can be journelled to a word processing file and then transferred into a spreadsheet to give a table such as Table 2, where the LSMs of ln(ODDR) can be translated back to marginal "LSM" estimates of Prob[Y=Hit] using formula (3). The spreadsheet data can then be used for tables of estimates and plots such as the one in Figure 4. Figure 4 suggests that although differences in hit performance between sights were generally not large as a function of range, performance of CANDB tended to fall-off faster with range. Since there is a physical explanation for such an increased fall-off, the plot proved to be helpful.

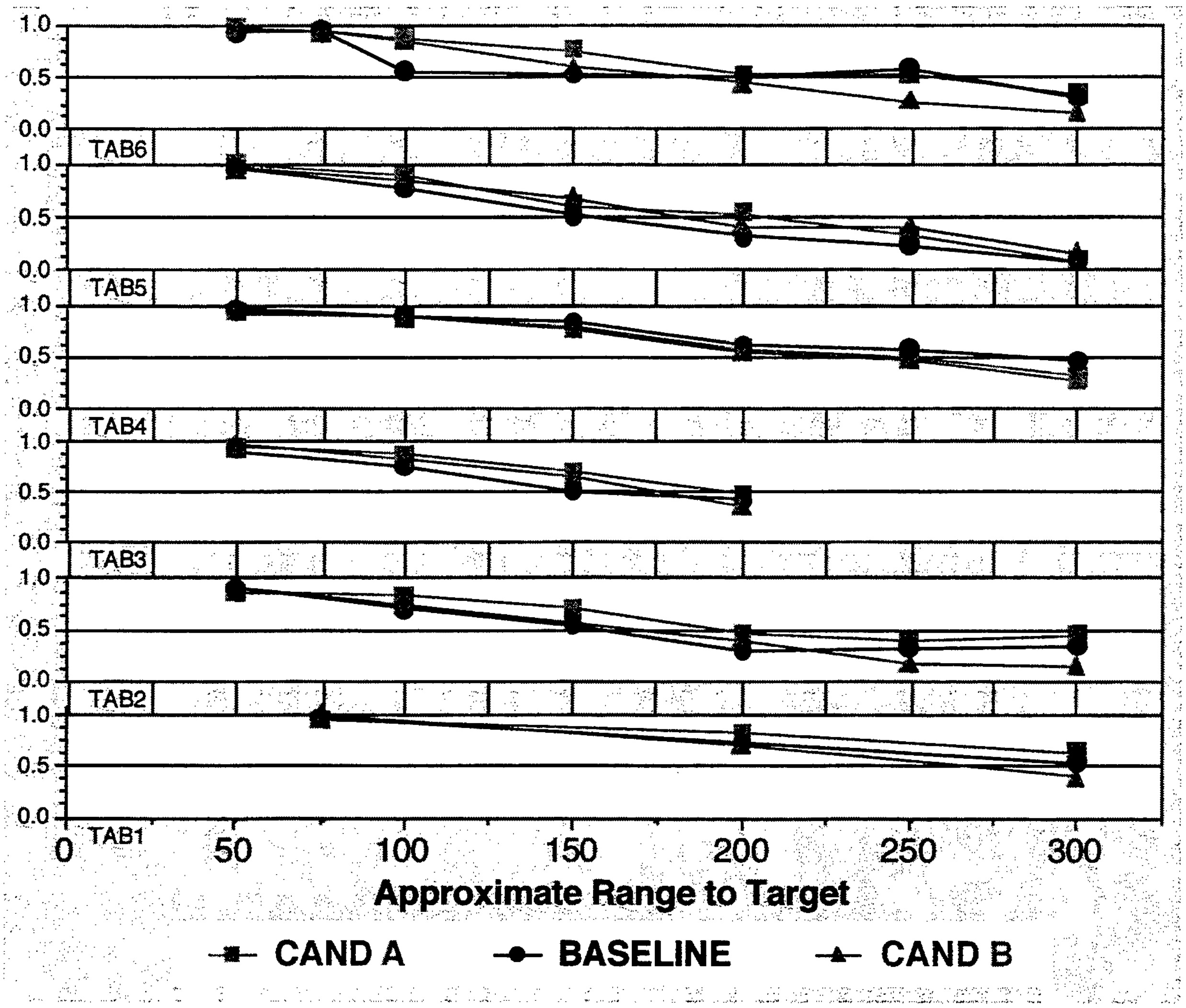


Figure 4. Estimated hit probabilities by RANGE for one WEAPON and all values of TABLE

SYNTHETIC PROBABILITIES

Figure 4 shows that graphical displays of effect estimates can be used to aid interpretation of logistic regression results in a manner similar to the way LSMs can be used to interpret ANOVA tables for continuous data. Although often helpful, displays such as Figure 4 suffer the problem common to all such displays of point estimates—there is no indication of spread and sample size. Once the vector λ of ln(ODDR) estimates is available, formula (3) can be used to for each of the row in the underlying contingency table to produce several types of interesting dot- and box-plots which help alleviate this problem. The key is that formula (3) produces for each of the 3890 rows in the underlying contingency table an estimate PROBHIT for Prob[Y=Hit] in that row which is based not only on PROPHIT=HITS/PRES for that row ("PRES" is the number of presentations—which varied from 1 to 11 with a mode of 5 or 6 having 840 presentations each) but also on many of the other 3889 rows via the model of Table 1. These PROBHIT estimates are analogous to the "predicted values" of OLS regression. But for count data they can

be regarded as "**synthetic proportions**" since they take the very coarse values for PROPHIT and smooth them via a complicated function (Figure 3) into much more continuous estimates suitable for graphical display. As a first example, the ordinary boxplots of synthetic proportions in Figure 5 produce a more satisfactory display than in the earlier Figure 2. With the amount of data in this example, however, even Figure 5 is too dense to be entirely pleasing. Multiway plots can spread out the data, alleviating this overly dense plotting. In fact, recent work at Bell Labs[2] has developed interesting tabular displays of graphical analyses called "trellis graphics" which are implemented in the newest versions of the "S" language. These displays permit flexible tabular display of multiway data, automatically smoothed or repackaged via a command language. Similar displays can be produced more clumsily outside S by carefully recoding horizontal and vertical plotting parameters in the data. In particular, the display of LSMs in Figure 4 can be replaced by the multiway display of dotplots in Figure 6. Compared to Figure 4, Figure 6 gives a deeper understanding of what is going on in the data, and it raises some disturbing questions since it is clear that hit performance is very closely grouped at short ranges.

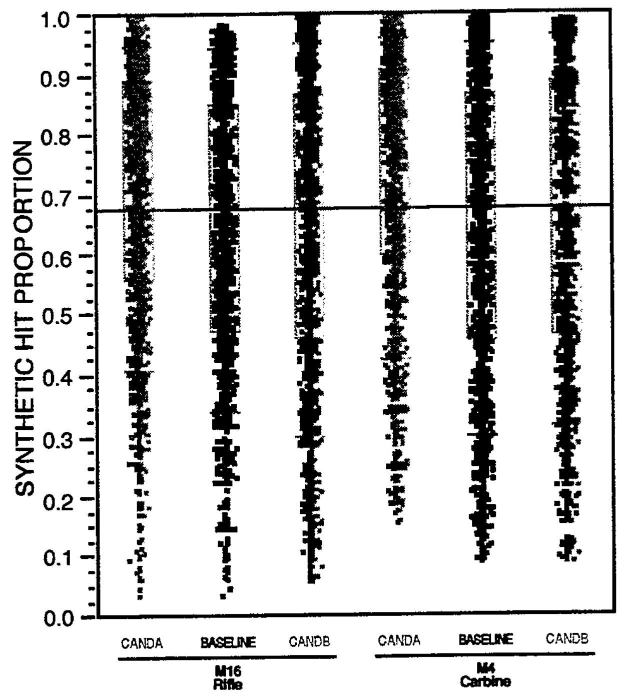


Figure 5. Boxplots of synthetic hit proportions (more helpful than Figure 2).

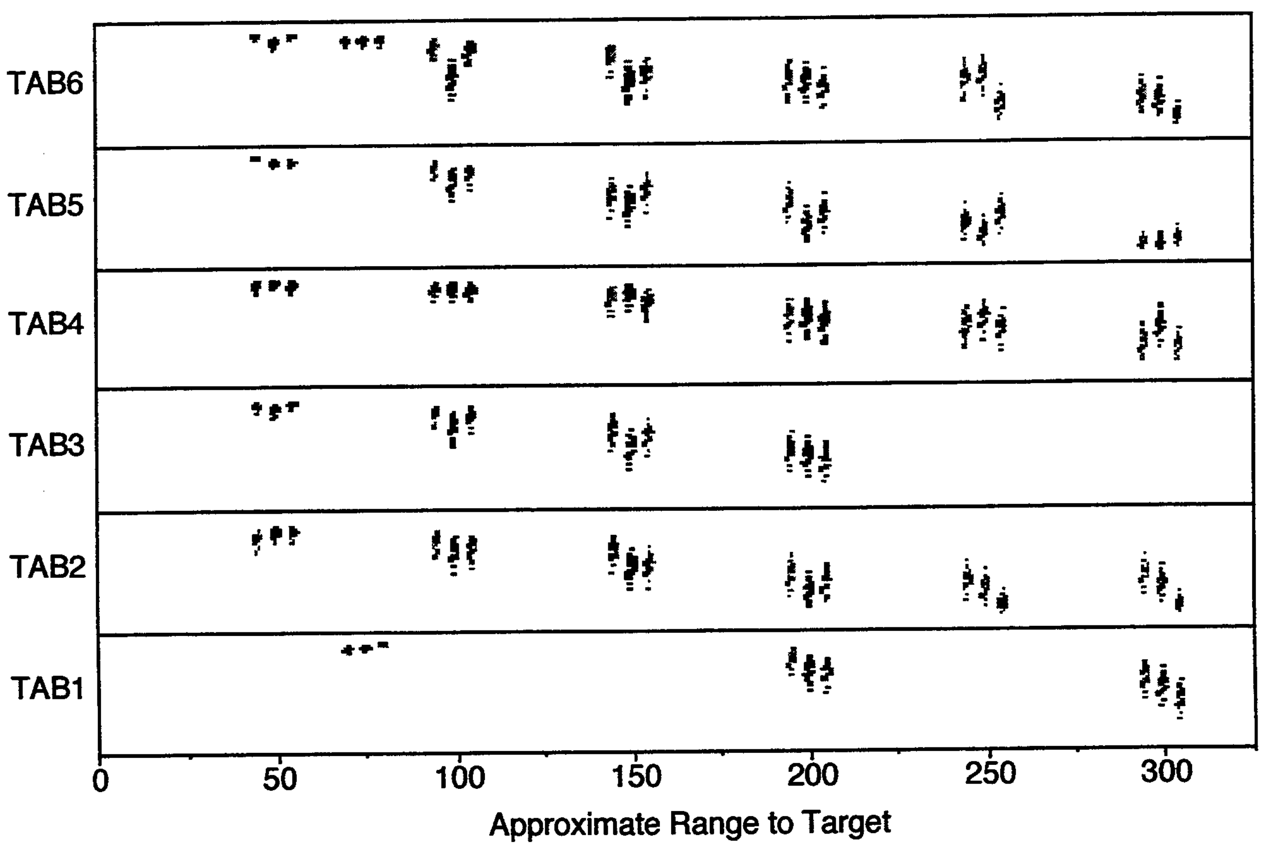


Figure 6. Synthetic hit proportions by range and firing table for one WEAPON.
(Left to right the jittered point clouds correspond to CANDA, BASELINE, and CANDB.)

RESIDUALS AND RESCALING

Although synthetic proportions permit graphical display of hit/miss data in a richer manner than the the true hit proportions do, they rely heavily on the underlying model. To assess the model dependence, some sort of residual plot is desirable, and a natural definition of residuals for plotting is

RESIDUAL = (HITS-PRES•PROBHIT)/SQRT(PRES•(1-PROBHIT)•PROBHIT). (4)

where "PRES" is the number of presentations. With this definition, the residual plot in Figure 7 is easy to obtain. Clearly something fishy is going on at short range and occasionally at long range. A few moments reflection is enough to guess the problem. With only a few presentations per cell and a very high probability of hit at close range, one would expect PROBHIT estimates to be very near 1 at short range so that residuals would be quite large (and negative) in any contingency table cell without perfect hit performance. Likewise, long range cases with relatively small hit probabilities could be expected to have some large positive residuals. The additive definition of residuals is not entirely satisfactory since the underlying model is multiplicative. However, the author does not know of a good way to formulate multiplicative residuals which accommodates the actual zeros and ones in the observed hit proportion data.

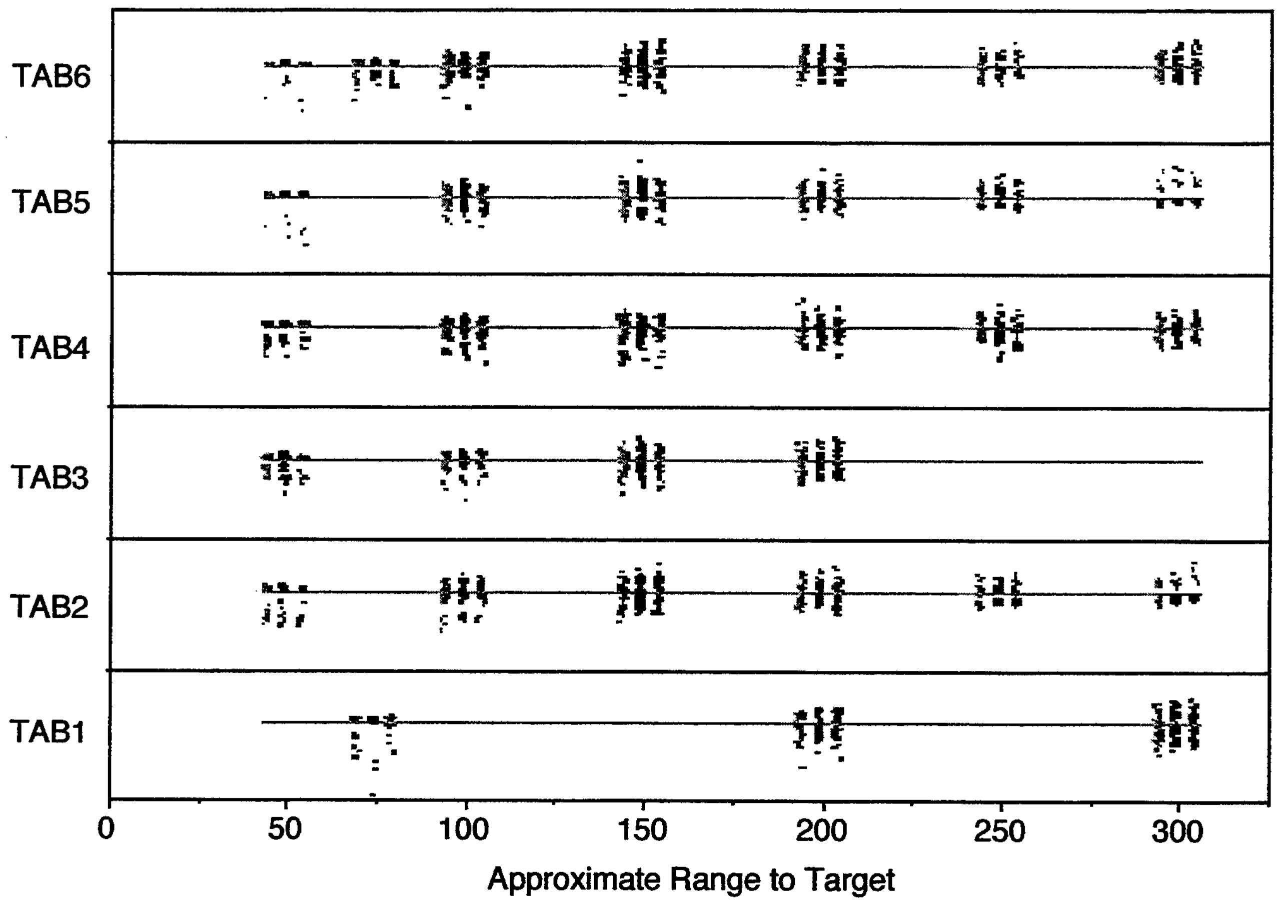


Figure 7. Residuals of synthetic hit proportions by range and firing table for one WEAPON.
(Left to right the jittered point clouds correspond to CANDA, BASELINE, and CANDB; the horizontal lines represent overall means.)

Plotting the original odds ratios (ODDR) on a log scale yields the annotated plot in Figure 8. Comparing Figure 8 with Figure 7 confirms that very high/low probabilities and small numbers of presentations (Figure 8) are associated with the large residuals in Figure 7. This plot also shows that there are two short-range cases which may

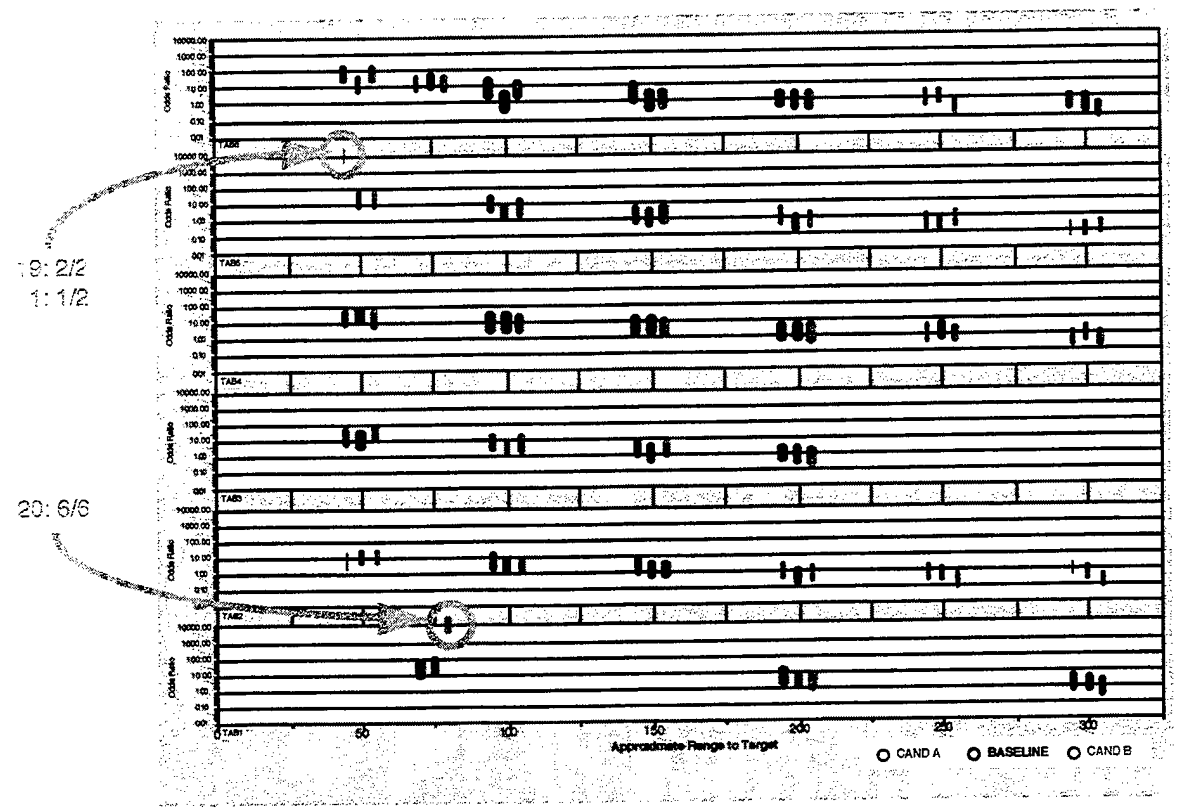


Figure 8. Rescaled plot of original odds ratio estimates on logarithmic scale (area of circles proportional to number of presentations).

have skewed the results and been influential in the logistic regression fit. One case involved presentations of only two targets where nineteen soldiers hit both while one soldier hit only one. The other involved presentations of six targets in which all twenty soldiers hit all six. The author does not know exactly how much these two apparently influential points affected the overall fit and significance statements presented in Table 1.

SUMMARY, CONCLUSIONS AND DIFFICULTIES

Clearly, graphical techniques can contribute to the understanding of count data analyzed via nominal logistic regression. Such techniques can provide substantial insight to both the model fit and the original data set. The notion of synthetic proportions helps a lot in providing helpful displays since it provides statistics which can be displayed and analyzed like continuous data. However, synthetic proportions rely heavily on the underlying model, they yield no really good residuals, and there is no clear path to influence diagnostics. Furthermore, the right plots are based on ln(ODDR), not PROBHIT (synthetic proportions). Finally, both the techniques and the graphics are borderline in both memory and processing time for most PCs; the word processing file in which this paper resides is 5.6MB in size, the number of objects in some graphics exceeded the 32K objects limitation in my graphics editing program and my printer had to be tricked into printing some pages. But these computational limitations are quickly disappearing. Despite the difficulties, the bottom line is that graphical techniques can be used effectively to provide insight to analysis of count data just as they are used effectively for continuous data.

REFERENCES

1. JMP Statistics and Graphics Guide. Cary, NC: SAS Institute, 1994.

2. Becker, Richard A., Cleveland, William S. and Shyu, Ming-Jen. "The Visual Design and Control of Trellis Display." Journal of Computational and Graphical Statistics, Volume 5, Number 2. pp. 123-155, 1996.

INTENTIONALLY LEFT BLANK.

An Intelligent Hierarchical Decision Architecture for Operational Test and Evaluation[1]

Major Suzanne M. Beers, USAF
Air Force Operational Test and Evaluation Center
HQ AFOTEC/CNP
8500 Gibson Blvd., SE
Kirtland AFB, NM 87117-5558
(505) 846 - 9929
Fax: (505) 846 - 9726
E-mail: beerss@afotec.af.mil

Dr. George J. Vachtsevanos
School of Electrical and Computer Engineering
Georgia Institute of Technology
777 Atlantic Drive
Atlanta, GA 30332-0250
(404) 894 - 6252
Fax: (404) 894 - 7583
E-mail: george.vachtsevanos@ece.gatech.edu

ABSTRACT

Since the inception of the Strategy-to-Task evaluation framework, originally suggested by RAND's Lt. Gen. Glenn A. Kent, the Operational Test and Evaluation community has been struggling with how to implement it. The top-down definition of the hierarchical structure linking high-level objectives and tasks to the functional performance that a system must demonstrate has been successfully accomplished. However, successful implementation of a methodology through which the functional performance level data gathered during testing can flow back up through the hierarchy: being aggregated and synthesized to provide truly meaningful information to the decision-maker has been elusive. This paper describes an Intelligent Hierarchical Decision Architecture that uses fuzzy set theory as well as the Dempster-Shafer Theory of Evidential Reasoning to take functional performance level data as input and provides a probabilistic bound on the system performance at the operational task level as output.

INTRODUCTION

The Strategy-to-Task evaluation framework, originally suggested by RAND's Lt. Gen. Glenn Kent (Kent & Simon, 1991), was eagerly adopted by the operational testing community as a means to link low-level functional performance information about a system, gathered during a testing effort, to high-level operational tasks and objectives that a system needs to be able to accomplish. Kent's hierarchical evaluation framework requires that high-level objectives be defined, then underlying objectives and operational tasks are outlined. Once the system's operational tasks are defined, the functional performance characteristics that a system must be able to meet to accomplish those operational tasks, are determined. This top-down definition of objectives to tasks to functional performance characteristics has been accomplished in many operational testing programs. What has been lacking is a methodology through which the functional performance level data gathered during the testing effort can be aggregated and synthesized to flow back up the strategy-to-task hierarchy to the operational task level, where it can provide meaningful information to the acquisition decision-maker.

Current analysis methods used by the Operational Test and Evaluation (OT&E) community are limited to standard statistical methods and a limited use of Modeling and Simulation (M&S). Although both have proved inadequate in providing information to the decision-maker at the operational task level, they continue to be used, and in fact, endorsed as the preferred analysis methods. The currently used statistical methods, such as, statistical hypothesis testing, analysis of variance, design of experiments, and non-parametric statistics offer a means of summarizing the information gathered during the testing efforts, but do not provide a method for extrapolating the data to higher information levels. Statistical model building techniques, such as, regression analysis and time series analysis provide a means to predict future performance once a model is built of a process, however, in most cases in the OT&E arena, sufficient data do not exist to build these models. M&S using the "legacy models" has been suggested as a means for answering questions at higher information levels, however, the M&S solution offers its own dilemmas. For example,

[1] Approved for public release: distribution is unlimited.

- In order for the models at the higher level (i.e., mission-level or campaign-level models) to run in a reasonable amount of time, many simplifications were made in their development. These simplifications preclude them from being used as detailed analysis tools.
- Each of the legacy models was developed by a different organization for a different purpose. There was no thought given to an architecture that would tie these models together until long after the models were already developed.
- Finally, the issue of verification, validation, and accreditation (VV&A) of these models is one that is just now beginning to receive attention. No systematic mechanisms or databases are readily available to allow the analyst to determine a model's applicability to the task at hand.

Other modeling techniques, such as Monte-Carlo simulation, can be employed to draw conclusions at the operational task level from the functional performance level data if transformations between the two information levels are known in functional form. However, in most cases, these functional transformations do not exist, thus, severely limiting the use of these methods. After an initial analysis of all of the statistical and analytical methods used in OT&E, the National Research Council affirmed the inadequacy of the current analysis methods, when they listed the four aspects of operational testing contributing to its difficulty and complexity (National Research Council, 1995):

- statistical methods meant for making one-at-a-time pass/fail decisions are inappropriate for OT&E decision-making problems
- OT involves realistic engagements where factors which cannot be controlled affect the testing outcome
- OT is expensive, thus, frequently the testing yields sparse data to support decision-making
- the incorporation of additional sources of relevant data poses methodological and organizational challenges.

So, we see that the OT&E community is faced with an analysis challenge: how to provide meaningful information to the acquisition decision-maker with currently available tools that are inadequate for the task. The OT&E community needs a methodology through which functional performance level data and other non-numerical information can be combined to help the decision-maker determine a system's task accomplishment capabilities. The method must be able to handle small data sample sizes, uncontrollable testing conditions, all relevant information regardless of its form, and not establish arbitrary pass/fail criteria. The *Intelligent Hierarchical Decision Architecture* has been developed to address this OT&E analysis void.

METHODOLOGY

This section describes a methodology through which low-level information is aggregated and synthesized to provide information at the operational task level using the Intelligent Hierarchical Decision Architecture, shown in Figure 1 (Beers 1996). The Intelligent Hierarchical Decision Architecture is composed of four components -- a Clustering Methodology which takes the raw test data and forms a fuzzy distribution, a Fuzzy Associative Memory which performs the transformation from the functional performance level to the operational task level, a Fuzzy Cognitive Map which adjusts the system performance measurement indicated by the testing effort for factors that could not be controlled or including in the testing, and an Aggregation Methodology which aggregates the system performance across the logical divisions of the system performance. First, we begin with a short description of fuzzy set theory, then describe each of the major components of the Intelligent Hierarchical Decision Architecture.

FUZZY SET THEORY BASICS

Throughout our formal mathematical education we are exposed to set theory. We learn in those early classes that an element is a member of a set or is not a member of a set -- black or white. Fuzzy set theory was introduced by Lofti Zadeh in 1965 to handle situations where an element can be a partial member of a set (Zadeh 1965) (Zadeh 1973). The degree of membership of an element within a fuzzy set is indicated by its membership function value, μ, a value in the range [0,1] with zero indicating no membership and unity indicating full membership. The values between zero and one are used to indicate partial membership of the element within the set. Consider the example of a man who is seven feet tall, clearly a member of the set of tall men, his membership function value with respect to the set would be unity, $\mu_{TALL} = 1.0$. On the other hand, a man who is 5'7" tall might be only a member of the set of tall men

only to a degree 0.5, $\mu_{TALL} = 0.5$. Finally, a man whose height was 3' clearly should not be considered a member of the set of tall men, thus his membership function value would be zero. The idea of this gradual transition from non-membership to full membership has found a use in many engineering applications, particularly systems control applications, where fuzzy set theory has improved the performance of such diverse equipment as subway trains, washing machines, and fault detection systems (McNeill & Frieberger 1990). It also offers a means for the testing community to consider system performance evaluations in a more realistic manner. With current analysis methods, the testing community must draw a line in the system performance space -- a hard and fast pass/fail criterion. However, the criterion is seldom that black and white. Why should an electronic combat system that causes a missile to miss an aircraft by 14'11" be considered a failure, while one that causes a miss distance of 15'1" be considered a success? Can we really justify that precision in our evaluation criteria, or would a gradual transition from bad to good performance be more realistic? The Intelligent Hierarchical Decision Architecture uses fuzzy set theory and fuzzy logic concepts throughout its processing to allow a more realistic and meaningful evaluation of the operational testing data. Now we turn to a discussion of the four stages of the hierarchical structure.

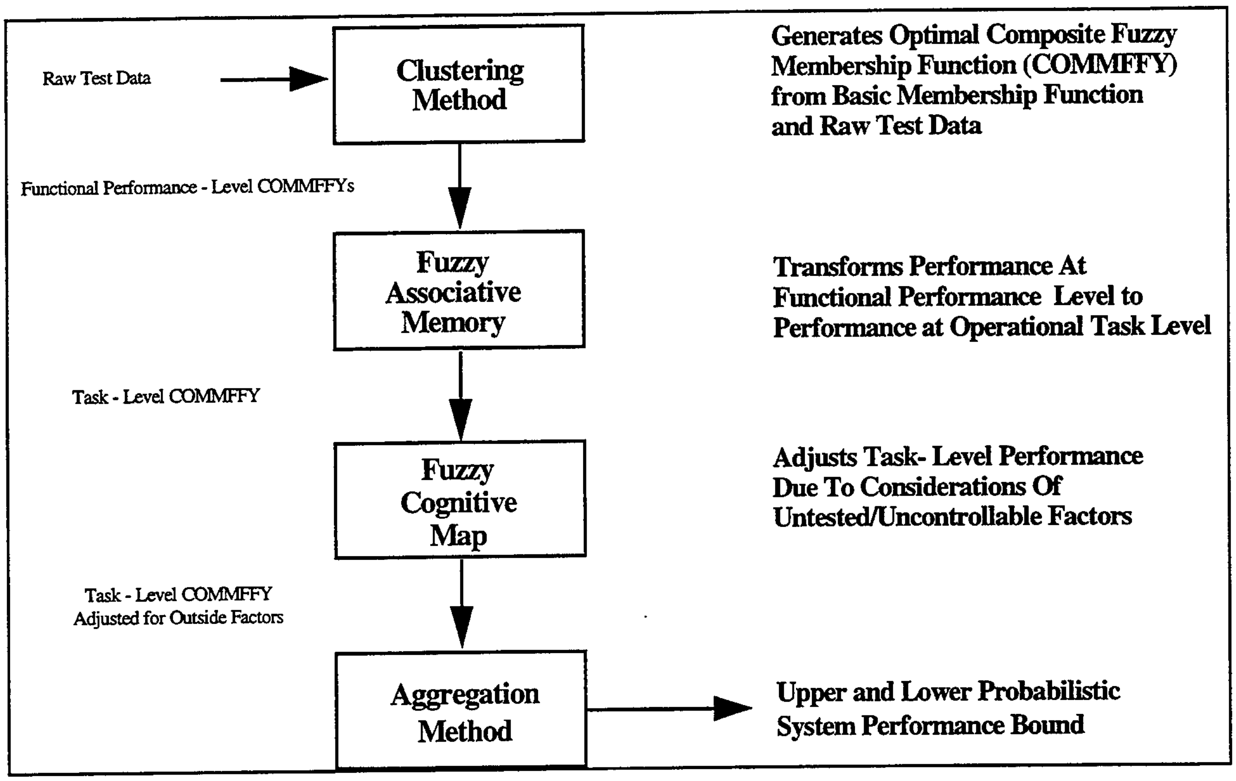


Figure 1 Intelligent Hierarchical Decision Architecture

STEP #1: CLUSTERING METHODOLOGY

The ***Clustering Methodology*** is the first stage in the Intelligent Hierarchical Decision Architecture. It takes the raw test data and forms it into a fuzzy set, which is called a *Composite Fuzzy Membership Function*, or COMMFFY. This COMMFFY, formed through the three step process described below, will be an optimal description of the original raw test data in fuzzy set form at the *Measure of Functional Performance* (MOFP) level. That is, each test measure for which data are gathered will have a COMMFFY built that will be used for subsequent processing within the Intelligent Hierarchical Decision Architecture.

The first step within the Clustering Methodology is to define fuzzy sets, called Basic Membership Functions, which will be the basis for constructing the COMMFFY. These fuzzy sets can be developed in one of two ways: through a fuzzy clustering method or through a heuristic approach. The fuzzy clustering method (Gath & Geva 1989) requires that enough data describing each Measure of Functional Performance be available to perform a fuzzy

clustering algorithm, which is frequently not the case during operational testing, so we will concentrate here on the heuristic approach. Using that approach, we look at all the possible values that a variable can take on, or its universe of discourse, and define fuzzy sets within the universe of discourse that adequately describe, in linguistic terms, those sets. For example, the triangular-shaped Basic Membership Functions shown in Figure 2 divide the universe of discourse into five equal segments with a 50% overlap. The linguistic tags LO, LOMED, MED, MEDHI, and HI describe the fuzzy sets and are used in subsequent stages to facilitate an intuitive understanding of the algorithmic processing.

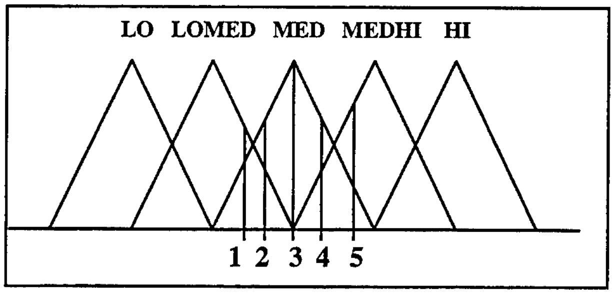


Figure 2 Sample Basic Membership Functions and Five Sample Data Points

Once the Basic Membership Functions have been defined, one of four *Compositional Methods* is used to form the Composite Fuzzy Membership Function, or COMMFFY. The Compositional Methods used to form the COMMFFY are Max-Max, Max-All, Min-Max, and Min-All. The four methods differ in how they apply the raw data points to the Basic Membership Functions. First, the inner operation is accomplished: either Max or All. Then, once the inner operation is accomplished, we look inside each Basic Membership Function to perform the outer operation: either Max or Min. Finally, the COMMFFY is formed by joining the components of the Basic Membership Functions derived from these two operations. The inner operation describes how each data point interacts with each Basic Membership Function, or fuzzy set. For example, with the xxx-Max operation, each data point activates only the fuzzy set where it is a maximum. In the sample shown in Figure 2, consider the data point labeled #1, it intersects both the MED fuzzy set and the LOMED fuzzy set. It is a maximum in the LOMED set. With data point #2, its maximum activation is in the MED set. Once the inner operation considers all the test data for a given measure, we turn to the outer operation. In this case, let's look at the Max-xxx operation. Now we look within each fuzzy set and find the maximum of all the activation levels generated by the inner operation. So in this example the maximum within the LOMED set was the activation level contributed by point #1, the maximum within the MED set was the activation level contributed by point #3, and so on. Once the inner and outer operations have been completed, the COMMFFY is formed by taking the maximum activation level within any Basic Membership Function for each member of the universe of discourse. Figure 3 shows the COMMFFY resulting from the Max-Max compositional method for the sample data points and Basic Membership Functions shown in Figure 2.

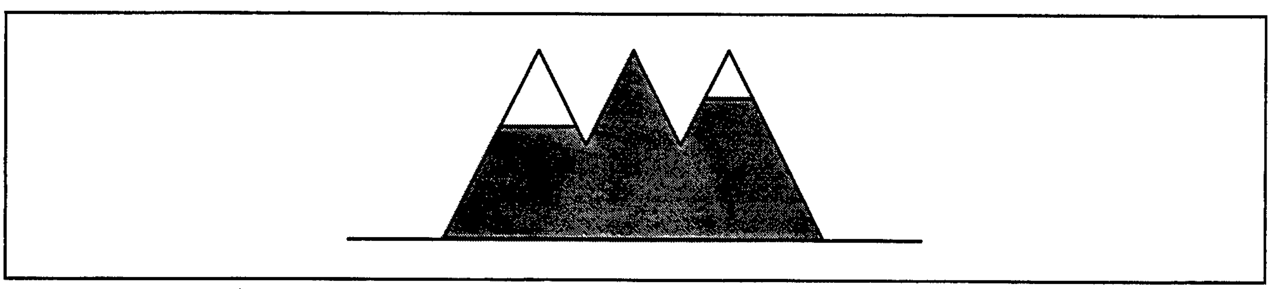


Figure 3 COMMFFY Resulting from Max-Max Compositional Method

The choice of which Compositional Method to use for a given data set is determined through an on-line optimization using a fuzzy/statistical similarity measure that was developed to relate the COMMFFY with a normal statistical distribution that would be generated from the same data. With the COMMFFY formed for each Measure

of Functional Performance, we now turn to the next stage in the Intelligent Hierarchical Decision Architecture, the *Fuzzy Associative Memory.*

STEP #2: FUZZY ASSOCIATIVE MEMORY

The second stage of the Intelligent Hierarchical Decision Architecture transforms the information at the functional performance level to information at the operational task level using a ***Fuzzy Associative Memory***, essentially a set of rules that relate the performance at the two levels in terms of fuzzy sets (Kosko 1992). Once the performance due to each of the functional performance measures has been transformed to the operational task level, the information is aggregated into a single COMMFFY at the operational task level using a modification of the Reduction Theorem (Wang & Vachtsevanos 1990).

The rules within the Fuzzy Associative Memory can initially be built using expert judgment, then subsequently updated as more information is gathered on the system-under-test's performance, through testing or modeling and simulation. The Fuzzy Associative Memory takes the form shown in Figure 4. Each of the boxes pictured in Figure 4 is a rule bank relating the fuzzy sets at the functional performance level with the fuzzy sets at the operational task level.

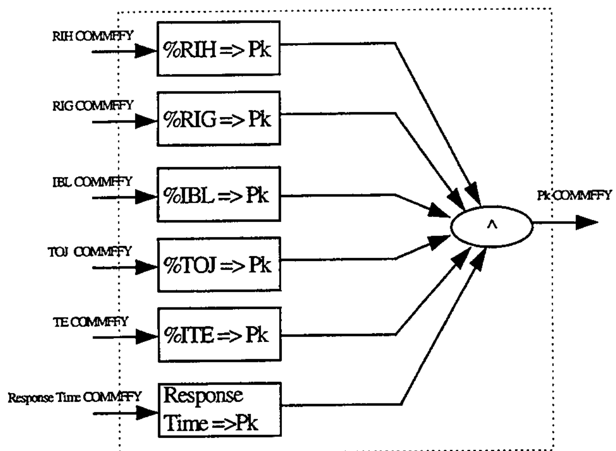


Figure 4 Intelligent Hierarchical Decision Architecture's Fuzzy Associative Memory Structure

The transformation from the functional performance level to the operational task level is accomplished using the Fuzzy Associative Memory as described above, yielding at the output of this second stage, a COMMFFY indicating the system's performance at the operational task-accomplishment level.

STEP #3: FUZZY COGNITIVE MAP

Frequently, during the performance of an operational test there are factors that cannot be included or controlled during the testing effort, yet are known to have an affect on the outcome of the system performance measure. To adjust the testing-derived system performance measurement for factors that could not be included or controlled in the testing effort, we use a ***Fuzzy Cognitive Map.***

A Fuzzy Cognitive Map is a figure indicating cause and effect relationships between factors, developed originally by Bart Kosko based upon the work done by Robert Axelrod (Axelrod 1976). Using the map, Kosko demonstrated that "what-if" questions could be answered by performing a series of matrix multiplication and thresholding operations on the matrix derived from the map (Kosko 1986). The Intelligent Hierarchical Decision Architecture uses Kosko's work as a foundation, and uses the map to adjust the system performance indicated by the test measurements for factors that could not be controlled or included during the testing effort. This adjustment is accomplished using the following steps:

- Define a Fuzzy Cognitive Map that relates the untestable and uncontrollable factors to the Measure of Task Accomplishment (MOTA) used during the testing effort. Figure 5 illustrates an FCM relating factors that could

affect the outcome during the testing of an electronic combat system. The linguistic tags define the degree and direction of the effects. For example, if the *Number of Threats in the Scenario* cannot be adequately represented (i.e., fewer threats on the test range than would be encountered in a wartime environment), the resulting *Reduction in P_k* that would be measured during the test would be ***some*** amount better than it would have been, therefore, a ***-some*** adjustment should be made to the measured performance.

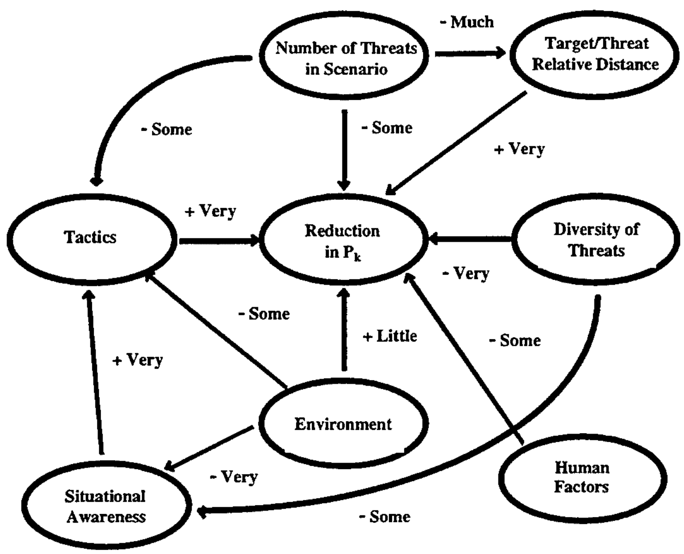


Figure 5 Sample Fuzzy Cognitive Map for an Electronic Combat System

- Looking at each concept within the map, define the possible paths from that concept to the task-accomplishment measure concept. For example, on the map shown in Figure 5, starting at *Number of Threats in Scenario* we can define a path directly to *Reduction in P_k*, and a path that goes through *Target/Threat Relative Distance* then to *Reduction in P_k*, etc. This path definition step can be simplified by using the matrix/vector multiplication to determine the limit cycles as described in (Kosko 1986). The activated concepts are those that need to be considered in the path definition process.

- Once all the possible paths from each concept to the central concept have been defined, find the minimum value of the linguistic tags associated with each path (this requires an importance ordering of the tags used to define the links, e.g., ***little* < *some* < *much* < *very***) ignoring the signs of the tags. Once the minimum value of each of the possible paths is defined, take the maximum value of the tags associated with each concept across all possible paths.

- Finally, rank order the most-negative to most-positive linguistic tags associated with all the concepts in the map. The most-negative tag will be used to adjust the task-level COMMFFY to indicate the worst-case system performance and the most-positive tag will be used to adjust the task-level COMMFFY to indicate the best-case system performance of the system.

- The adjustment is carried out using the following adjustment formulae:

$$\mu_{PosAdj} = \mu_{old}^{1/k}$$ *Best Case Adjustment*

$$\mu_{NegAdj} = \mu_{old}^{k}$$ *Worst-Case Adjustment*

where the value of k is chosen to provide an adequate adjustment to the fuzzy distribution values and μ represents the fuzzy membership function values.

STEP #4: AGGREGATION METHODOLOGY

The first three stages of the Intelligent Hierarchical Decision Architecture are carried out for each logical division of the system-under-test's performance. For example, for the testing of an electronic combat system, the first three stages would be carried out for each threat system that the electronic combat system is tested against. The final stage aggregates the system performance across the logical divisions, providing the final result, a probabilistic bound on the system performance at the operational task level.

The aggregation is carried out using Dempster's Rule of Combination taken from the Dempster-Shafer Theory of Evidential Reasoning (Shafer 1976). Using this method, each of the adjusted, task-level COMMFFYs for the best-case system performance are combined to form a best-case probabilistic bound; and each of the worst-case COMMFFYs are combined to form a worst-case probabilistic bound. These two probabilistic bounds, along with a measure of the Degree of Certainty associated with each possible hypothesis, are provided to the decision-maker as the outcome of the operational testing effort. The basic steps of the aggregation method are as follows.

- The maximum degree of membership within each of the original Basic Membership Functions is defined from the COMMFFYs. For each logical division of the system-under-test's performance, possible hypotheses sets are defined by taking alpha-level cuts of the fuzzy set defined from the Basic Membership Function values. Subtracting, subsequent values of the alpha-level cuts gives the Dempster-Shafer basic probability assignment value for each hypothesis (Yen 1990). The basic probability assignment is the amount of evidence that is pointing to that hypothesis being true.

- The evidence associated with each logical division of the system-under-test performance is then combined two-by-two with other division's evidence using the intersection tableau method (Gordon & Shortliffe 1985), which provides a logical application of Dempster's Rule of Combination. Given two pieces of evidence that provide information on the hypothesis Ψ denoted $m_1(\Psi)$ and $m_2(\Psi)$, the combined basic probability assignment is denoted $m_{12}(\Psi)$ and is given by:

$$m_{12}(\Psi) = \frac{\sum_{A \cap B = \Psi} m_1(A) m_2(B)}{1-K}$$

where K is a normalization factor and is given by:

$$K = \sum_{A \cap B = \varnothing} m_1(A) m_2(B)$$

- The belief function, defined as the lower probability bound on a hypothesis, and the plausibility function, defined as the upper probability bound on the hypothesis are calculated using the formulae shown below, where $m(A)$ is the basic probability assignment value associated with hypothesis A (deKorvin & Shipley 1993).

$$Bel(B) = \sum_{A \subset B} m(A)$$

$$Pl(B) = \sum_{A \cap B \neq \varnothing} m(A) = 1 - Bel(\overline{B})$$

- Finally, the Degree of Certainty, is a value in the range [-1,+1] that indicates the amount of evidence pointing to the hypothesis as opposed to the amount of evidence pointing to contradicting hypotheses (Kim 1992). A value of +1 for the degree of certainty indicates that all the evidence is pointing to the hypothesis and none is pointing to contradicting hypotheses, a value of -1 indicates that all the evidence is pointing to the contradicting hypotheses, and a value of zero indicates total ignorance, in that equal amounts of evidence are pointing to the hypothesis and the contradicting hypotheses. The degree of certainty is calculated as

$$DOC(X) = m(X) - Bel(\overline{X})$$

With the belief and plausibility functions and the degree of certainty calculated, the decision-maker is provided with the final probabilistic bound of the system performance at the task-accomplishment level. The final result is given in the form:

> *The best-case system performance is* ***linguistic tag*** *(where the tag is associated with the basic membership function(s) representing the most likely hypothesis) with probability range [0.xxxx,0.yyyy](where 0.xxxx is the belief function value and 0.yyyy is the plausibility function value). The degree of certainty associated with this statement is zz% (where zz is the degree of certainty).*

EXAMPLE

To illustrate the methodology described in the previous section, a brief example will be given here. Consider an aircraft-mounted jammer system, that when tested, has six Measures of Functional Performance (MOFPs) and is tested against four separate threat systems. The decision-maker is interested in determining the system's ability to reduce the probability of kill of the aircraft carrying the jammer. The evaluation framework for the system, called Jammer-X, is shown in Figure 6.

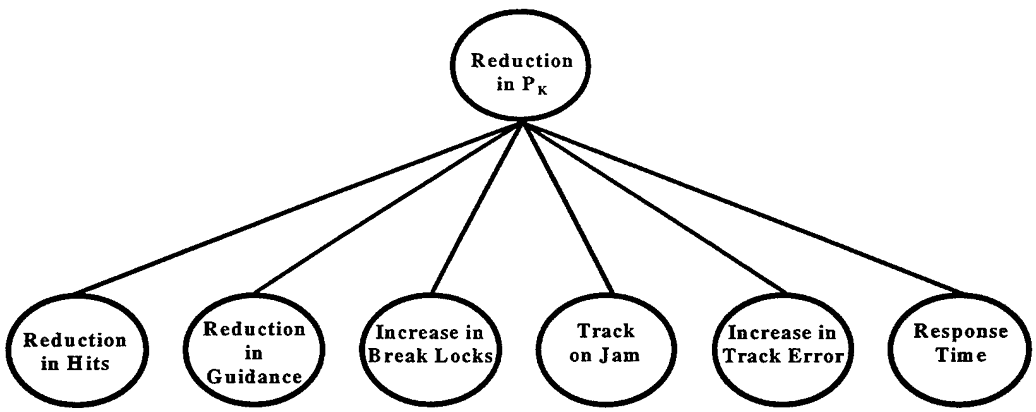


Figure 6 Jammer-X Evaluation Framework

During the OT&E, data are gathered on each of the functional performance measures, and current analysis methods provide the decision-maker with 24 pass/fail results at that level requiring that he draw high-level conclusions from this low-level information. The Intelligent Hierarchical Decision Architecture will be used to form a probabilistic bound on the system's *Reduction in* P_k capabilities, based upon the measurements taken on the aspects of the system's technical performance shown in Figure 6. The first step is to define the Basic Membership Functions and apply the test data to them to form a Composite Fuzzy Membership Function, or COMMFFY, for each MOFP/threat combination. The Basic Membership Functions chosen for this example are triangular-shaped, with a 50% overlap, as shown in Figure 7.

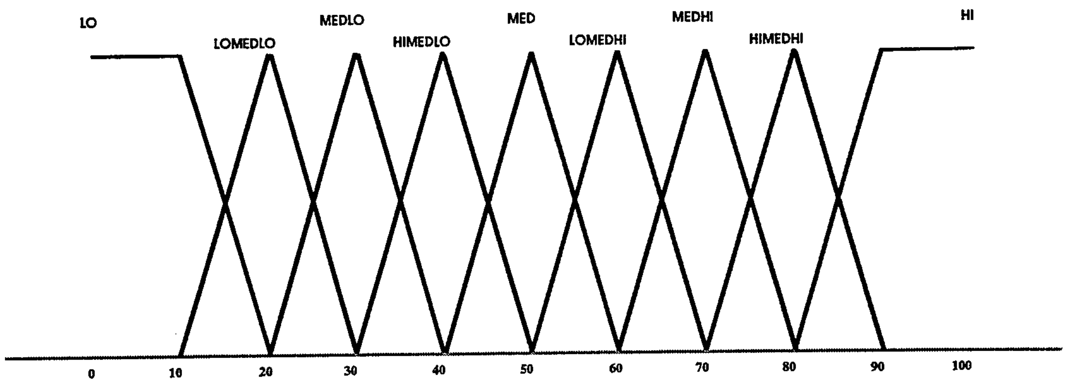
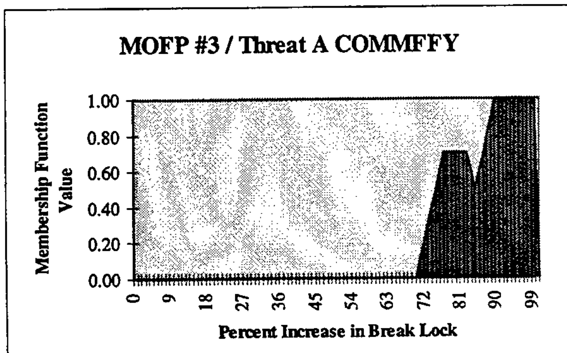


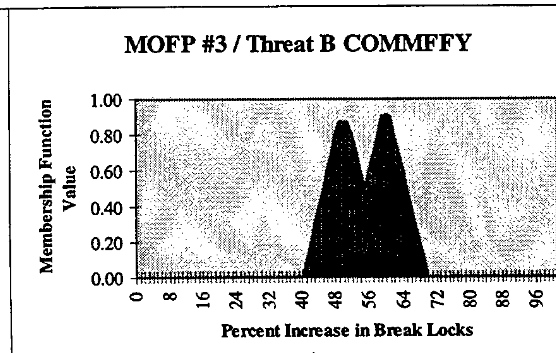
Figure 7 Basic Membership Functions

Using the basic membership functions shown in Figure 7 as the foundation, and applying the data given as an example of the test measurements taken on one of the MOFPs, shown in Table 1, the COMMFFYs illustrated in Figure 8 result.

	Percent Increase in Break Locks			
Run Number	**Threat A**	**Threat B**	**Threat C**	**Threat D**
1	76.65	57.17	43.64	23.78
2	89.77	53.44	47.77	35.93
3	90.92	55.46	35.93	16.65
4	90.47	62.48	48.65	48.65
5	98.66	62.95	31.90	52.97
6	94.78	51.33	49.40	68.85
7	91.38	59.11	38.41	59.11
8	94.10	53.73	46.44	46.15
9	76.15	52.97	30.63	77.02
10	77.02	51.63	35.10	76.15

Table 1 Raw Test Data Collected for MOFP #3, Percent Increase in Break Locks





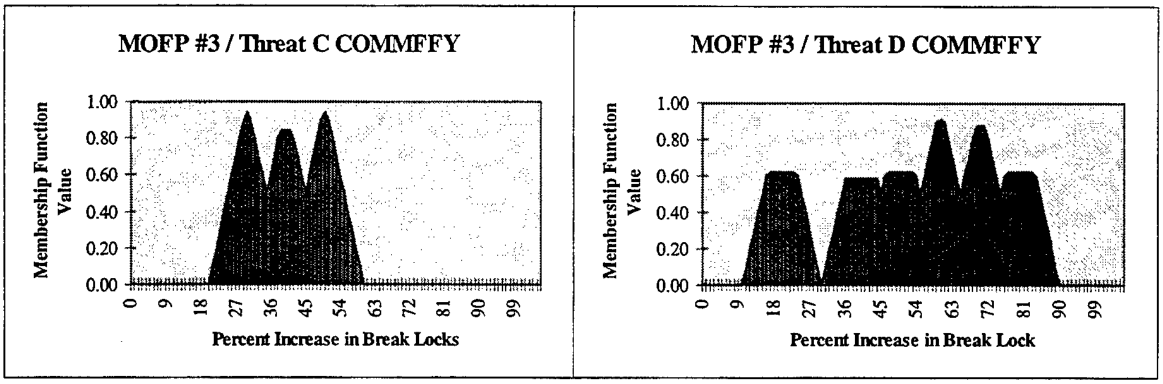


Figure 8 Functional Performance Level COMMFFYs for MOFP #3

The COMMFFYs illustrated in Figure 8 are four of the 24 that would be formed in the first stage of the hierarchy's processing for this example. Once all the functional performance level COMMFFYs have been formed, the

Fuzzy Associative Memory is used to transform these fuzzy distributions to COMMFFYs at the task accomplishment level, as described in Step #2 of the methodology section. Each COMMFFY formed at the Measure of Task Accomplishment (MOTA) level results from the aggregation of the six functional performance level COMMFFYs for that threat. The resulting MOTA-level COMMFFYs for this example are shown in Figure 9.

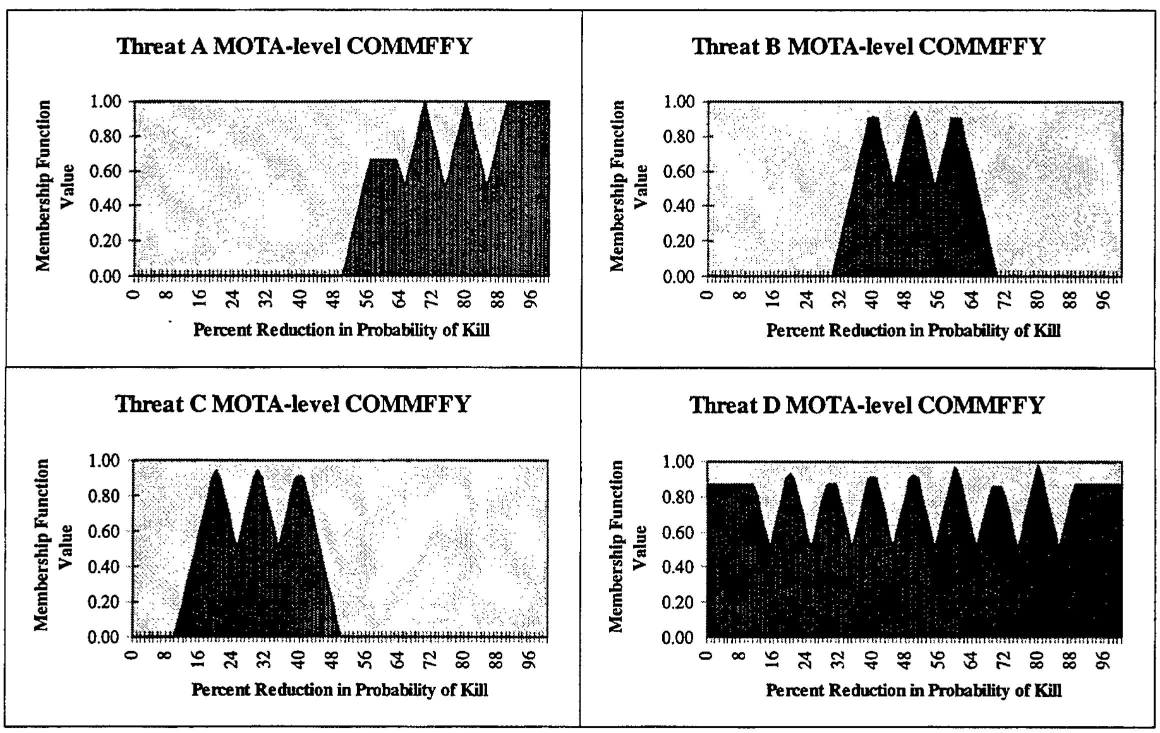


Figure 9 Task Accomplishment Level COMMFFYs for Jammer-X

The COMMFFYs shown in Figure 9 represent the task-level system performance demonstrated during the testing effort. In most cases, the testing effort cannot include or control all the factors known to affect system performance. Therefore, in the third stage of the Intelligent Hierarchical Decision Architecture these COMMFFYs are adjusted for the effects of those factors, as described in Step #3 of the methodology section. Using the Fuzzy Cognitive Map

shown in Figure 5, the best-case adjustment is ***+very*** and the worst-case adjustment is ***-very***. If an adjustment factor of 2.0 is used in association with the linguistic tag ***very***, then the COMMFFYs resulting from this adjustment for the Threat B performance, look like those shown in Figure 10.

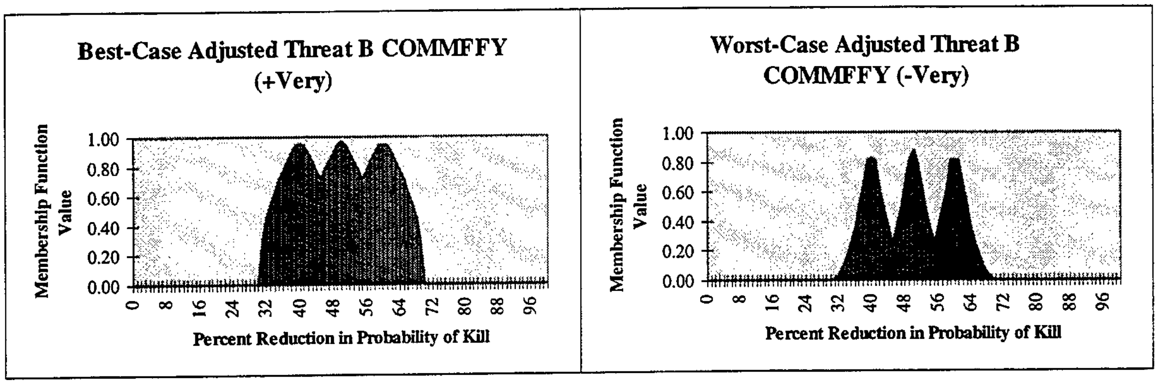


Figure 10 Adjusted Task Accomplishment Level Performance Against Threat B

Finally, in order to provide a single, probabilistic system performance bound to the decision-maker, the Dempster-Shafer theory is used, as described in Step #4 of the methodology section. The information provided to the decision-maker would be as shown below.

*The Jammer-X's Best-Case Performance is **LoMedHi** (the Basic Membership Function centered at 60% Reduction in P_k) with probability range [0.9621, 0.9925]. The degree of certainty associated with this statement is 92.42%.*

*The Jammer-X's Worst-Case Performance is **LoMedHi** (the Basic Membership Function centered at 60% Reduction in P_k) with probability range [0.7443, 0.8545]. The degree of certainty associated with this statement is 48.86%.*

CONCLUSION

Since the inception of the Strategy-to-Task evaluation framework, the operational test community has struggled with a way of taking the low-level test data that is generated during testing events or through modeling and simulation, and use it to provide information to the acquisition decision-maker that is meaningful to the decisions being made.

Current analysis methods used by the community are limited to standard statistical methods which provide a means for summarizing the information, but do not readily provide a means for extrapolating the gathered information to higher information levels where it is meaningful for the decision being made. Modeling and simulation efforts, such as Monte-Carlo simulation, could be used, but do not allow for the consideration of qualitative information or allow a realistic approach to the analysis that includes gradual transitions from good-to-bad system performance. The Intelligent Hierarchical Decision Architecture described here can be used to take the low-level functional performance data generated during the testing effort and synthesize and aggregate it into a probabilistic system bound at the operational task level. In addition to simply considering the information gathered during the testing, it allows a method through which non-testable or non-controllable factors can be considered. It allows the consideration of qualitative as well as quantitative information and is not constrained by sample size requirements, as are current statistical methods. The methodology allows smooth transitions from good-to-bad system performance, yet yields a definitive statement, in probabilistic terms, on the system's capabilities, as a final output. With this methodology, the operational test community can more adequately provide the information that the acquisition decision-making community expects.

REFERENCES

Axelrod, R., *Structure of Decision: The Cognitive Maps of Political Elites*, Princeton, NJ: Princeton University Press, 1976.

Beers, S.M., *An Intelligent Hierarchical Decision Architecture for Operational Test and Evaluation*, Ph.D. Dissertation, Georgia Institute of Technology, May 1996.

Committee on National Statistics of the National Research Council, *Statistical Methods for Testing and Evaluating Defense Systems*, Interim Report, Washington D.C.: National Academy Press, 1995.

de Korvin A. and Shipley, M.F., "A Dempster-Shafer-Based Approach to Compromise Decision Making with Multiattributes Applied to Product Selection," *IEEE Transactions on Engineering Management*, Vol. 40, No. 1, pp. 60-67, February 1993.

Gath, I. and Geva, A.B., "Unsupervised Optimal Fuzzy Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 7, July 1989.

Gordon, J. and Shortliffe, E.H., "A Method for Managing Evidential Reasoning in a Hierarchical Hypothesis Space," *Artificial Intelligence*, Vol. 26, No. 3, pp. 323-357, July 1985.

Kent, G.A. and Simons, W.E., *A Framework for Enhancing Operational Capabilities*, RAND Project Air Force Report, No. R-4043-AF, Santa Monica: RAND, 1991.

Kim, I., *A Hybrid Analytical/Intelligent Methodology for Sensor Fusion*, Ph.D. Dissertation, Georgia Institute of Technology, December 1992.

Kosko, B., "Fuzzy Cognitive Maps," *International Journal of Man-Machine Studies*, Vol. 24, pp. 65-75, January 1986.

Kosko, B., *Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence*, Englewood Cliffs, NJ: Prentice Hall, 1992.

McNeill, D. and Frieberger, P., *Fuzzy Logic*, 1990.

Shafer, G., *Mathematical Theory of Evidence*, Princeton, NJ: Princeton University Press, 1976.

Wang, B.H. and Vachtsevanos, G.J., "Fuzzy Associative Memories: Identification and Control of Complex Systems," *Proceedings of the Fifth International Symposium on Intelligent Control*, pp. 910-915, 1990.

Yen, J., "Generalizing the Dempster-Shafer Theory to Fuzzy Sets," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 20, No. 3, pp. 559-570, May/June 1990.

Zadeh, L.A., "Fuzzy Sets," *Information and Control*, Vol. 8, pp. 338-353, 1965.

Zadeh, L.A., "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-3, No. 3, pp. 28-44, January 1973.

AN APPROACH TO GENERATING BAYESIAN PROBABILITY OF BELIEF IN MISSILE P_k

Michael B. Dewitz and Paul W. Ellner
U.S. Army Materiel Systems Analysis Activity
ATTN: AMXSY-EP
392 Hopkins Road
Aberdeen Proving Ground, MD 21005

ABSTRACT

Army decision makers are forced to rely heavily on the results of simulations when making programmatic decisions about developmental systems. Quantifying a measure of assurance associated with achieving a specified level of P_k has been an ongoing problem in the Army community.

The U.S. Army Materiel Systems Analysis Activity (AMSAA) has developed methodology based on Bayesian analysis that quantifies the probability of belief associated with the P_k output from a simulation model. The approach is to quantify the distribution of uncertainty in the input parameters to the simulation model based on available test data. This uncertainty is used to generate a distribution of belief regarding the output P_k of the simulation model. The generated P_k belief distribution gives the Bayesian probability that the specified level of P_k has been achieved. This paper describes the methodology developed by AMSAA and discusses an example that demonstrates the applicability of the methodology.

INTRODUCTION

When Army missile system development programs reach a milestone decision point, estimates of true system performance are compared to required performance (expressed by probability of kill (P_k)) as a means of determining whether or not the program should continue. Because the true system performance is unknown, equally important as the estimate of true performance is a measure of assurance that the true performance exceeds the requirement. When the data used to develop the estimates of system performance come from system level testing of actual hardware, that assurance is often expressed in terms of confidence bounds using classical statistical analysis techniques. The statistical confidence is dependent on the number of system flight tests. In today's environment, funding available for testing Army missile systems is decreasing. As these systems become more complex, testing also becomes more complicated and expensive. Many developmental programs are conducting more component and subsystem testing to quantify the performance parameters that define overall system performance and then executing complex simulations to relate these performance parameters to P_k. There are fewer tests of the entire system where system P_k can be measured directly. This presents the challenge of determining a suitable, quantifiable, measure of assurance that true system performance exceeds the requirement when data come from simulations and a wide variety of test sources.

In May 1996, the U.S. Army Materiel Systems Analysis Activity (AMSAA) proposed methodology to the office of the Deputy Undersecretary of the Army for Operations Research (DUSA-OR) which provides a quantifiable measure of assurance that true system performance exceeds a stated goal by using Bayesian techniques. The DUSA-OR office asked that an example using the proposed methodology be conducted to evaluate its merits. The proposed methodology and the results of the example conducted for the DUSA-OR follow in this report. Throughout this report, the measure of system performance that will be discussed is P_k.

METHODOLOGY

Before discussing the methodology, it is important to understand the basic approach being advocated. Essentially, when decision makers want to know what assurance the system developers have in their performance estimates (expressed in terms of P_k), they want to know what assurance there is that the true, but unknown, P_k exceeds some goal (typically an operational requirement). The approach proposed in this report is to quantify that assurance through the Bayesian measure of belief probability. This concept is illustrated in Figure 1.

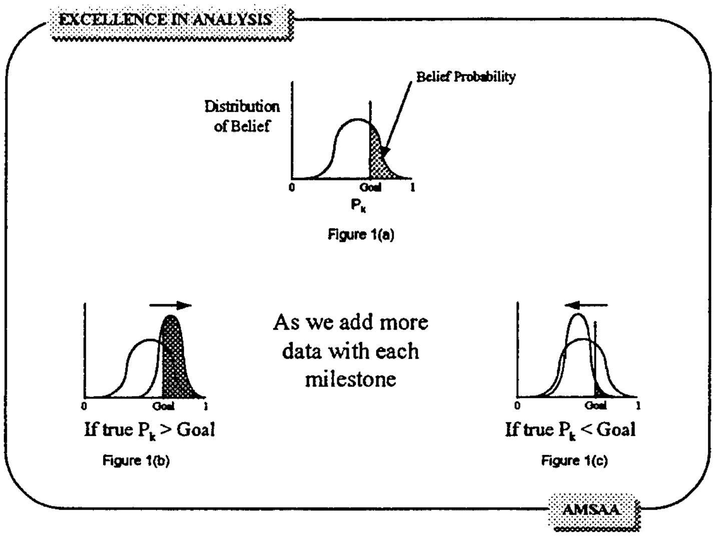


Figure 1. Bayesian Distribution of Belief Curves.

Overall system performance, as measured by P_k, is determined by the outcome of a series of more elementary processes. These elementary processes are governed by performance parameters that define the missile system at the component level. As an example, one of the elementary processes a missile system must execute is target track. One of the performance parameters that govern target track is the angular measurement accuracy of the missile seeker. When the relationship between the performance parameters that define a missile system and the resultant P_k cannot be expressed analytically, one can use a simulation. Many of the performance parameters being simulated are stochastic and are represented by their distribution parameters in the simulation. By randomly selecting from the distributions defining the performance parameter inputs and running the simulation in a Monte Carlo fashion, an estimate of system performance ($\hat{P}_k$) is generated. Although typically expressed as a point estimate, in truth there is some distribution of belief regarding P_k (shown in Figure 1(a)) that defines the entire realm of possibilities for P_k. This distribution of belief is the result of the uncertainty one has regarding the distribution parameters that define the performance parameter inputs (assuming a correct simulation model). Because there is uncertainty, the distribution parameters can have a range of values. A different $\hat{P}_k$ is generated each time different values for the distribution parameters are used.

Once the distribution of belief is quantified, the belief probability that P_k lies in any given interval can be determined. If one selects a goal value for P_k, then the area under the distribution of belief curve that lies to the right of the goal (assuming a scale that increases from left to right) is the belief probability that the true, but unknown, missile P_k exceeds the goal. As more knowledge is gained regarding the system, the uncertainty in the distribution parameters decreases which results in a narrower distribution of belief regarding P_k (Figures 1(b) and 1(c)). If the true P_k exceeds the goal chosen (Figure 1(b)), one would expect the distribution of belief to shift to the right. The result would be that a larger percentage of the curve will exceed the goal, and the belief probability that the true P_k exceeds the goal will increase. If the true P_k does not exceed the goal (Figure 1(c)), the distribution will shift to the left, a smaller percentage of the curve will exceed the goal, and the belief probability that the true P_k exceeds the goal will decrease. Even if the distribution of belief regarding P_k does not shift left or right, the belief

probability that the true system performance exceeds the goal will change as more knowledge is gained because the shape of the distribution will change.

In the current process for generating performance data using P_k simulations, the uncertainty in the distribution parameters that define the distributions of the performance parameters is not captured. For each engagement point in space and for each target of interest, the simulation is executed a fixed number of times using the *nominal* distribution parameters associated with each of the performance parameters. Each simulation replication results in either a kill or no kill. Dividing the total number of kills by the total number of replications yields a point estimate of system performance ($\hat{P}_k$) for a given engagement point and threat. The methodology proposed by AMSAA captures the uncertainty in the distribution parameters of the critical performance parameters in the P_k simulation thereby generating a distribution of belief regarding missile P_k.

Stated generally, the methodology is to first identify the set of critical performance parameter inputs (X_1, X_2, ..., X_k) that have a significant impact on the simulation output. As an illustration, assume the X_i are independent normally distributed random variables with means θ_i and standard deviations σ_i (i=1, ..., k). The next step is to characterize the distribution of uncertainty in either θ_i, σ_i, or both. As an example, assume each θ_i is known with certainty, but each σ_i has uncertainty about it that is represented by some probability density function, $\Omega(\sigma_i)$. Each distribution of uncertainty is determined by the body of knowledge (test data, physics, engineering judgment, requirements, etc.) about that particular performance parameter input (X_i) at a given point in time and is called the prior distribution of uncertainty for σ_i.

Once the distributions of uncertainty are characterized, they can be introduced into the P_k simulation. The process for introducing the uncertainty in the σ_i is to randomly select a value for each σ_i (where i takes values from 1 to k) from each of the distributions of uncertainty (i.e., each $\Omega(\sigma_i)$). This determines the distribution parameters defining each X_i in the simulation. The next step is to run sufficient replications of the simulation where each replication uses the set of values for the random variables X_i drawn from the their respective distributions to generate $\hat{P}_k$, an estimate of P_k (where $\hat{P}_k$ equals the number of kills achieved divided by the total number of replications as discussed above). Note that $\hat{P}_k$ is generated using fixed θ_i and σ_i (i=1,...,k). Next, draw new random values for the σ_i (i=1,...,k) from the distributions of uncertainty and repeat the process. The result will be a different $\hat{P}_k$. Repeat the entire process a sufficient number of times to generate a histogram of $\hat{P}_k$ outcomes. This histogram is termed the estimated prior distribution of belief for P_k. If a goal value for P_k is selected, the percent of area to the right of the goal is the estimated belief probability that the true P_k exceeds the goal. As the system under development progresses, more data will be gathered for each of the X_i performance parameter inputs. This new body of knowledge is used to update the distributions of uncertainty of each σ_i. The updated distribution of uncertainty is termed the posterior distribution. The entire process is repeated using the current posterior distribution of uncertainty for each σ_i to develop the corresponding posterior distribution of belief regarding P_k.

There is one important assumption that must be satisfied when selecting the critical performance parameter inputs for this analysis. Because the analysis is based on reducing the uncertainty about the true, but unknown, distribution parameters θ_i and σ_i, it is imperative that the true, but unknown, θ_i and σ_i do not change. The implication is that the X_i selected for this analysis must have true, but unknown, θ_i and σ_i that do not vary from test-to-test, do not vary throughout a given test event, are not affected by changes made to the P_k simulation, and do not change as the system matures. The limitation of this assumption is that some of the critical performance parameters may not be included in the analysis. Additionally, it is desirable that the X_i performance parameter inputs be independent. If there is dependence, it must be explicitly accounted for.

In addition to providing a quantifiable measure of assurance that the true P_k exceeds some goal, there is another powerful benefit of the approach. The distribution of belief regarding P_k is determined by the uncertainty one has regarding the distribution parameters that define the performance parameter distributional inputs. This uncertainty is updated as new information is gathered on the performance parameters through testing. One can therefore optimize a test strategy that focuses on reducing the uncertainty in those distribution parameters that have the greatest contribution to the diffuseness of the distribution of belief regarding P_k.

THE EXAMPLE

To exercise the methodology, a P_k simulation existing at AMSAA was used. To simplify the process for the example requested by the DUSA-OR, the uncertainty in the distribution parameters of only one performance parameter input was considered. The variable selected for this example is normally distributed. Based on the present body of knowledge, the mean (θ) is known to be 0.0. The uncertainty lies with the standard deviation (σ).

If no data exist to form a distribution of uncertainty regarding the standard deviation, one can assume a noninformative prior distribution of uncertainty. The distribution, $\Omega(\sigma)$, used for the standard deviation assuming a noninformative prior is $1/\sigma$ for $\sigma > 0$. Although the noninformative prior distribution of uncertainty is a starting point, it is not a proper density function and cannot be used to create a distribution of belief for P_k. This distribution of uncertainty must first be updated with data. The data in Table 1 were used to accomplish this update. The analyses were conducted using only the first two data points to update the distribution of uncertainty regarding the standard deviation and then repeated using all nineteen data points. The purpose for doing it twice is to show how the distribution of uncertainty for the standard deviation, the distribution of belief regarding P_k, and the belief probability that the true P_k exceeds some goal change as more information is obtained.

Table 1. Data Used to Update Distribution of Uncertainty in the Standard Deviation

i	x_i		i	x_i
1	0.417		11	0.365
2	-0.417		12	-0.340
3	0.355		13	0.365
4	-0.355		14	-0.360
5	0.350		15	0.360
6	-0.335		16	-0.365
7	0.340		17	0.365
8	-0.350		18	-0.365
9	0.355		19	0.370
10	-0.350			

The data (x_1, x_2,..., x_{19}) in Table 1 denoted by the vector, $\underline{x}$, were used to update $\Omega(\sigma)$ and generate a distribution of uncertainty for the standard deviation given the data. The updated distribution of uncertainty for σ, given $\theta=0$, and $\underline{x}$ (denoted by $\Omega(\sigma/\theta=0,\underline{x})$) takes the form;

$$\Omega(\sigma / \theta = 0, \underline{x}) = K \bullet L(\underline{x}; \theta = 0, \sigma) \bullet \Omega(\sigma) \quad \text{where:} \tag{3.1}$$

K = normalizing constant to ensure $\int_0^\infty \Omega(\sigma / \theta = 0, \underline{x}) d\sigma = 1$

$L(\underline{x};\theta=0,\sigma)$ = likelihood function associated with $\underline{x}$, from the normal distribution with mean 0 and standard deviation, σ. By definition,

$$L(\underline{x}; \theta = 0, \sigma) = \prod_{i=1}^{n} \eta(x_i; \theta = 0, \sigma) \quad \text{where:} \tag{3.2}$$

$\prod_{i=1}^{n}$ = product from i=1 to i=n;

$\eta(x_i;\theta=0,\sigma)$ = probability density function for x_i given $\theta=0$ and σ

For a normally distributed random variable,

$$\eta(x_i;\theta=0,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\bullet e^{-\frac{1}{2}\left(\frac{x_i-\theta}{\sigma}\right)^2}=\frac{1}{\sigma\sqrt{2\pi}}\bullet e^{-\frac{1}{2}\left(\frac{x_i}{\sigma}\right)^2}\qquad \text{since } \theta=0.$$

Substituting into equation (3.2) yields,

$$L(\underline{x};\theta=0,\sigma)=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\bullet e^{-\frac{1}{2}\left(\frac{x_i}{\sigma}\right)^2}=\frac{1}{\sigma^n(2\pi)^{\frac{n}{2}}}\bullet e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2}$$

Substituting into equation (3.1) yields,

$$\Omega(\sigma/\theta=0,\underline{x})=\frac{K}{\sigma^{(n+1)}(2\pi)^{\frac{n}{2}}}\bullet e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2}\tag{3.3}$$

By definition, the integral of the probability density function, $\Omega(\sigma/\theta=0,\underline{x})$, from zero to infinity is one. Utilizing this definition and solving for K yields,

$$K=\left[\frac{1}{(2\pi)^{\frac{n}{2}}}\int_0^\infty \sigma^{-(n+1)}\bullet e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2}\,d\sigma\right]^{-1}\tag{3.4}$$

Equation (3.4) can be solved explicitly by making the following substitution,

Let $y=\sigma^{-2}\sum_{i=1}^{n}x_i^2$ then $dy=-2\sigma^{-3}\sum_{i=1}^{n}x_i^2\,d\sigma$

Note $y\to\infty$ as $\sigma\to 0$ and
$y\to 0$ as $\sigma\to\infty$

Substituting into equation (3.4) yields,

$$K=\left[\frac{1}{2(2\pi)^{\frac{n}{2}}\left(\sum_{i=1}^{n}x_i^2\right)^{\frac{n}{2}}}\int_0^\infty y^{\frac{n}{2}-1}e^{-\frac{y}{2}}dy\right]^{-1}=\left[\frac{1}{2(2\pi)^{\frac{n}{2}}\left(\sum_{i=1}^{n}x_i^2\right)^{\frac{n}{2}}}\frac{\Gamma\left(\frac{n}{2}\right)2^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)2^{\frac{n}{2}}}\int_0^\infty y^{\frac{n}{2}-1}e^{-\frac{y}{2}}dy\right]^{-1}\tag{3.5}$$

The probability density function for the chi-square distribution with n degrees of freedom is,

$$f(x)=\frac{1}{\Gamma\left(\frac{n}{2}\right)2^{\frac{n}{2}}}y^{\frac{n}{2}-1}e^{-\frac{y}{2}}$$

Since the integral of the gamma density function from zero to infinity is one, substituting into equation (3.5) yields,

$$K = \left[\frac{1}{(2\pi)^{\frac{n}{2}}} \bullet \frac{\Gamma\left(\frac{n}{2}\right) 2^{\frac{n}{2}-1}}{\left(\sum_{i=1}^{n} x_i^2\right)^{\frac{n}{2}}} \right]^{-1} = \frac{2\left(\pi \sum_{i=1}^{n} x_i^2\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} \qquad (3.6)$$

Substituting the expression (3.6) back into equation (3.3) yields,

$$\Omega(\sigma / \theta = 0, \underline{x}) = \frac{2\left(\pi \sum_{i=1}^{n} x_i^2\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} \bullet \frac{1}{\sigma^{n+1} (2\pi)^{\frac{n}{2}}} \bullet e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2}$$

or,

$$\Omega(\sigma / \theta = 0, \underline{x}) = A \bullet \sigma^{-(n+1)} \bullet e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2} \quad \text{where,} \quad A = \frac{\left(\sum_{i=1}^{n} x_i^2\right)^{\frac{n}{2}}}{2^{\frac{n}{2}-1} \Gamma\left(\frac{n}{2}\right)} \qquad (3.7)$$

The posterior distribution of uncertainty for s is now defined. The next step in the process is to choose random samples of the standard deviation from this distribution of uncertainty and use them in the P_k simulation to generate the distribution of belief regarding P_k. To sample the standard deviation, one needs to randomly select a value, u, from a uniformly distributed random variable, U[0,1], and solve the following expression for σ_0:

$$u = \text{Prob}(\sigma \le \sigma_0) = \int_0^{\sigma_0} \Omega(\omega / \theta = 0, \underline{x}) d\omega$$

Note that by substituting in equation (3.7),

$$\text{Prob}(\sigma \le \sigma_0) = \int_0^{\sigma_0} A \bullet \omega^{-(n+1)} \bullet e^{-\frac{1}{2\omega^2} \sum_{i=1}^{n} x_i^2} d\omega \qquad (3.8)$$

Equation (3.8) can be solved explicitly by making the following substitution:

$$y = \omega^{-2} \sum_{i=1}^{n} x_i^2 . \quad \text{Then } dy = -2\omega^{-3} \sum_{i=1}^{n} x_i^2 d\omega ,$$

$$y \to \infty \text{ as } \omega \to 0, \text{ and}$$

$$y \to \sum_{i=1}^{n} \left(\frac{x_i}{\sigma_0}\right)^2 \text{ as } \omega \to \sigma_0.$$

Substituting into equation (3.8) yields,

$$\text{Prob}(\sigma \le \sigma_0) = \frac{A}{2\left(\sum_{i=1}^{n} x_i^2\right)^{\frac{n}{2}}} \bullet \int_{\sum_{i=1}^{n}\left(\frac{x_i}{\sigma_0}\right)^2}^{\infty} y^{\frac{n}{2}-1} e^{-\frac{y}{2}} dy = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_{\sum_{i=1}^{n}\left(\frac{x_i}{\sigma_0}\right)^2}^{\infty} y^{\frac{n}{2}-1} e^{-\frac{y}{2}} dy = \overline{G_{\chi_n^2}}\left[\sum_{i=1}^{n}\left(\frac{x_i}{\sigma_0}\right)^2\right]$$

where $\overline{G_{\chi_n^2}}(\chi^2) = 1 - G_{\chi_n^2}(\chi^2)$, and $G_{\chi_n^2}$ is the cumulative distribution function of a χ^2 random variable with n degrees of freedom. Since u= Prob($\sigma \le \sigma_0$),

$$G_{\chi_n^2}\left[\sum_{i=1}^{n}\left(\frac{x_i}{\sigma_0}\right)^2\right] = 1 - u$$

Equivalently,

$$\sum_{i=1}^{n}\left(\frac{x_i}{\sigma_0}\right)^2 = \chi^2_{n,1-u} \tag{3.9}$$

where $\chi^2_{n,1-u}$ denotes the 1-u percentile point of a chi-squared random variable with n degrees of freedom. Solving equation (3.9) for σ_0 yields,

$$\sigma_0 = \left(\frac{\sum_{i=1}^{n} x_i^2}{\chi^2_{n,1-u}}\right)^{\frac{1}{2}}$$

Since U is uniform from 0 to 1 if, and only if, 1-U is uniform from 0 to 1, we have

$$\sigma = \left(\frac{\sum_{i=1}^{n} x_i^2}{\chi^2_{n,U}}\right)^{\frac{1}{2}} \tag{3.10}$$

Equation (3.10) expresses σ, treated as a random variable due to uncertainty regarding its true value, as a function of the uniform random variable on [0, 1] and the test data. Note σ only depends on the test data through the values of n and $\sum_{i=1}^{n} x_i^2$.

From equation (3.7) or (3.10), it is clear that as more data are acquired that the distribution of uncertainty for the standard deviation changes. From the data in Table 1, equation (3.10) simplifies to;

$$\sigma = \frac{0.5897}{\left(\chi^2_{2,u}\right)^{\frac{1}{2}}} \text{ for n=2,} \qquad \text{and } \sigma = \frac{1.5808}{\left(\chi^2_{19,u}\right)^{\frac{1}{2}}} \text{ for n=19.}$$

For a given sample size n, one can randomly select a value u from U [0, 1], determine the corresponding χ^2_n percentile, and compute a random value for σ. The χ^2_n cumulative distribution functions that relate u to χ^2_n percentiles, $\chi^2_{n,u}$, for n=2 and n=19 are shown in Figures 2 and 3, respectively.

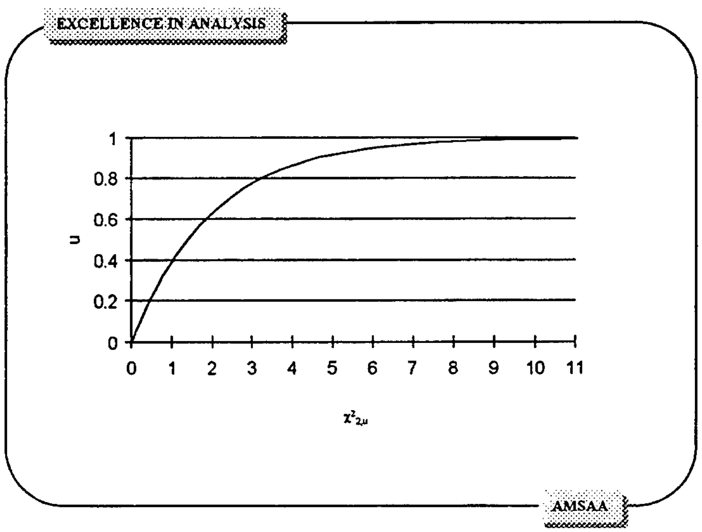


Figure 2. Chi-squared Cumulative Distribution Function for Two Degrees of Freedom.

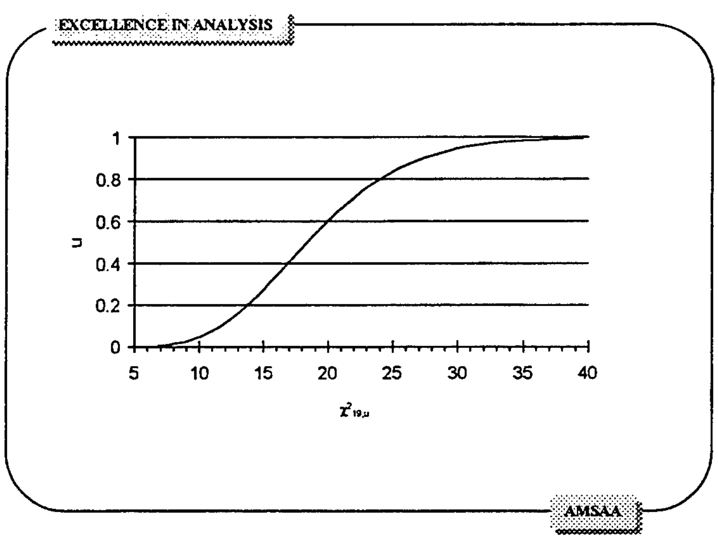


Figure 3. Chi-squared Cumulative Distribution Function for Nineteen Degrees of Freedom.

Figure 4 shows the probability density functions when two and 19 data points are used to update the uncertainty regarding the standard deviation. The sample standard deviation when two data points are considered is 0.590. From Figure 4, one can see that there is a high probability that the standard deviation will take on a value near 0.590. The distribution of uncertainty is also very diffuse and takes on a wide variety of values with significant probability. When all 19 data points are considered, the sample standard deviation is 0.372. Again, the distribution of uncertainty is highly weighted in that area. Now, however, the distribution is across a much narrower range of values for the standard deviation. As more data about the performance parameter are collected, the uncertainty in the distribution parameter that defines the performance parameter decreases. The methodology proposed in this report quantifies the distribution of uncertainty and allows for it to be introduced into the P_k

simulation.

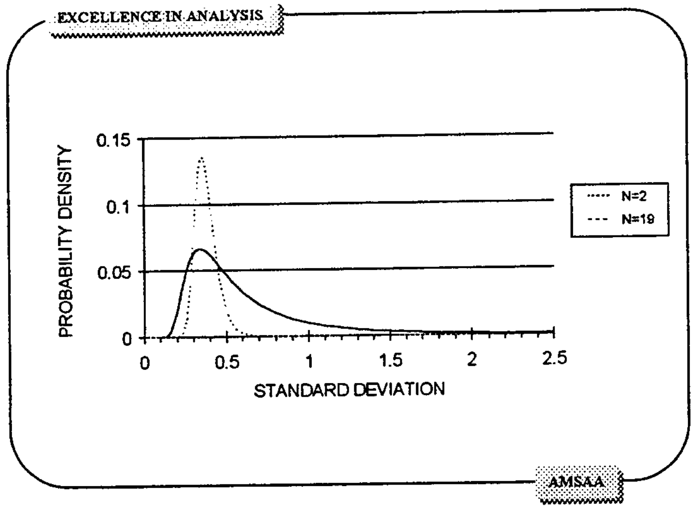


Figure 4. Posterior Distributions of Uncertainty Regarding Performance Parameter Standard Deviation Given Two and Nineteen Data Points.

To create the distribution of belief regarding P_k, 600 values for u were selected randomly from U[0,1]. These values for u were then used to determine 600 values of σ. Methodology development is ongoing to provide a means of determining the required number of random number draws. One method would be to compare the drawn distribution of uncertainty with the analytical expression for it. When the difference between the two is below some acceptable threshold, one would no longer select another value. The cumulative distribution functions for the uncertainty regarding the standard deviation using two and 19 data points are shown in Figure 5. The agreement between the analytical functions and those generated by randomly drawing 600 values is quite good (maximum difference less than 5 percent) indicating that 600 draws is sufficient to adequately characterize the distribution of uncertainty for this example.

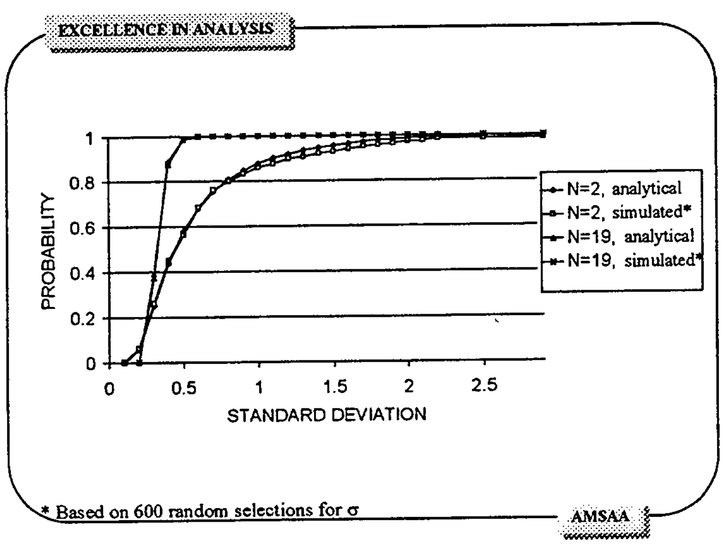


Figure 5. Comparison of the Analytical and Simulated Cumulative Distribution Functions for the Uncertainty in the Performance Parameter Standard Deviation Using Two and Nineteen Data Points.

To relate the distribution of uncertainty in the performance parameter standard deviation to P_k, one must exercise the simulation. Figure 6 shows the sensitivity of $\hat{P}_k$ to variations in the performance parameter standard deviation. The points in the figure are the estimates of P_k generated using a 25 replication Monte Carlo set of runs and a given value for the standard deviation. The line in the figure is a third order polynomial curve fit to the data. For this example, the $\hat{P}_k$ values corresponding to the randomly drawn sample of 600 σ's were used when relating the performance parameter standard deviation to P_k. The methodology can be exercised two ways. One way is to execute the simulation for each value of the standard deviation drawn from the distribution of uncertainty. The other way is to execute the simulation using a sufficient number of different standard deviations to construct the relationship between the performance parameter standard deviation and P_k. One could then use this relationship to compute P_k values that correspond to randomly selected standard deviations instead of executing the simulation. AMSAA is currently investigating the efficiencies of each approach as part of the follow-on effort to this report.

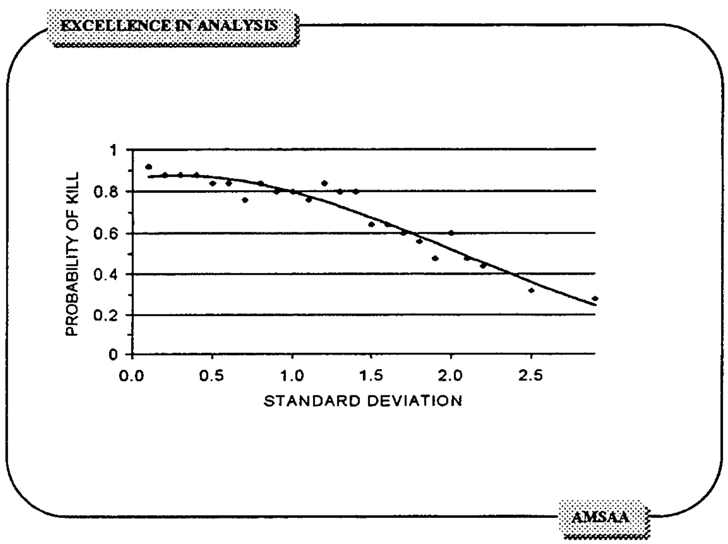


Figure 6. Sensitivity of $\hat{P}_k$ Estimates Using a 25 Replication Monte Carlo Set to Variations in Performance Parameter Standard Deviation.

Creating an estimated distribution of belief regarding P_k is simply a matter of combining the probability density for the standard deviation in Figure 4 with the P_k versus standard deviation estimated relationship displayed in Figure 6. The estimated distributions of belief regarding P_k when two and 19 data points are used to update the distribution of uncertainty regarding s are shown in Figure 7. Note the diffuse nature of the distribution of belief when two data points are used relative to the distribution of belief when 19 data points are used. Because the uncertainty in the performance parameter standard deviation decreases with more data, the resulting distribution of belief becomes more focused about a single value.

Recall from Figure 1 that the belief probability that the true, but unknown, P_k exceeds some goal is simply the area under the distribution of belief curve that lies to the right of that goal. Figure 8 shows the estimated belief probability that the true P_k exceeds any goal P_k. The sample standard deviation for the two data points in this example is 0.590. From Figure 6, the $\hat{P}_k$ associated with that standard deviation is 0.84. This is typically the only information given to decision makers. Using the information in Figure 8, one can also give the decision maker some assurance that the true system performance exceeds the estimate of performance by associating an estimated belief probability of 0.72 with the point estimate of 0.84. After 19 data points have been collected, the sample standard deviation decreases to 0.372. The missile $\hat{P}_k$ associated with that standard deviation is 0.88 (from Figure 6). The estimated belief probability associated with that missile $\hat{P}_k$ is 0.85. For any goal P_k value, the additional data points gathered resulted in an increase in the estimated belief probability that the true P_k exceeds that goal. Additionally, with the information provided by the 19 data points, the true P_k exceeds 0.84 (i.e., the estimated belief probability is 1.0) and the true P_k does not exceed 0.92 (i.e., the estimated belief probability is 0.0).

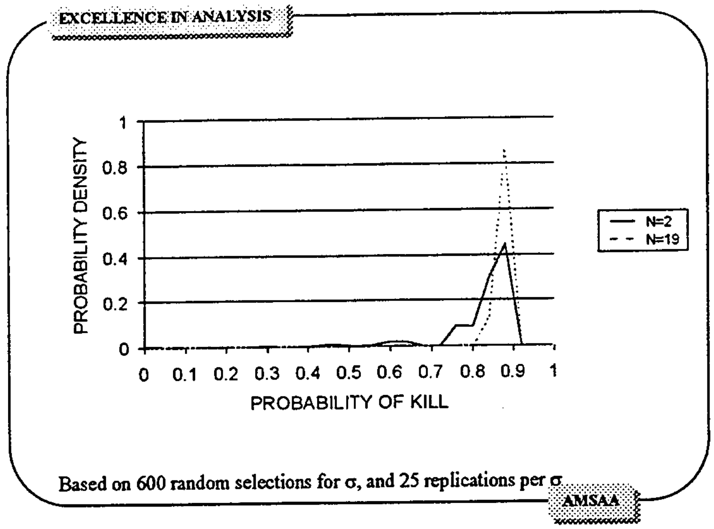


Figure 7. Estimated Posterior Distributions of Belief Regarding P_k Created Using Two and Nineteen Data Points.

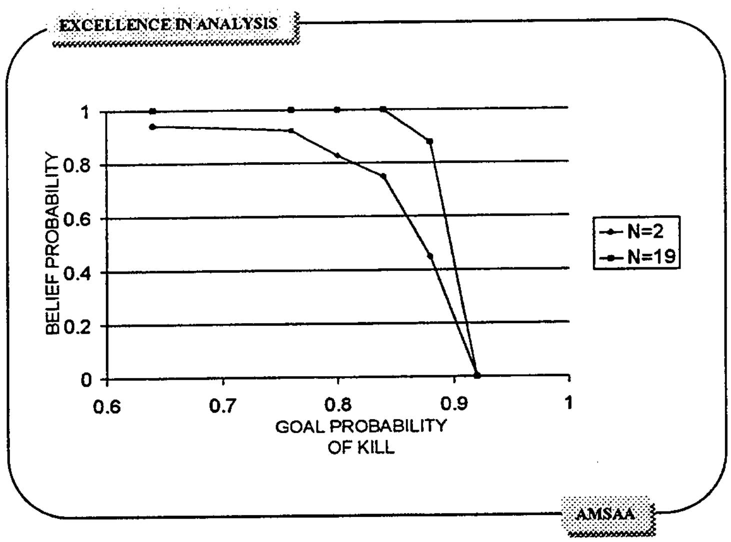


Figure 8. Estimated Belief Probability that the True P_k Exceeds any Given Goal Using Two and Nineteen Data Points.

To illustrate how this information could be used by programmatic decision makers, consider that the system is required to achieve a P_k of 0.8. Early in the life cycle of the program, say Milestone I, the decision maker may be willing to accept a moderate amount of risk. This manifests itself in allowing the program to proceed even if there is a low (say 0.7) estimated belief probability that the true missile P_k exceeds the requirement. By using the test data generated during the program's life cycle to update the uncertainty about the distribution parameters of key performance parameters, the estimated distribution of belief regarding P_k will change. As the program matures, one would expect the decision maker to demand less risk, so the program must demonstrate a higher estimated belief probability that the true P_k exceeds the goal before it would be allowed to proceed. If the system is exceeding its goal P_k, a larger portion of the updated estimated distribution of belief curve regarding P_k will lie to the right of the goal (0.8 in this discussion) provided our simulation estimates of P_k are sufficiently accurate, thus demonstrating a higher estimated belief probability that the true system performance exceeds the goal.

As discussed earlier, this methodology can be used to determine the optimum allocation of fixed resources for the collection of data to minimize risk or to minimize data required to demonstrate a given level of risk. The

distribution of belief regarding P_k is related to the uncertainty in the distribution parameters which is reduced by acquiring test data. Therefore, a test strategy can be optimized to focus on reducing the uncertainty in those distribution parameters that contribute most to the diffuseness of the distribution of belief regarding P_k. This methodology depicts the cause and effect relationship between distribution parameter uncertainty and distribution of belief regarding P_k thereby providing valuable information in the development of a test strategy.

CONCLUSIONS

For some time, the Army community has been trying to determine a way of providing assurance to decision makers that true overall system performance meets requirements when estimates of system performance come from simulations. A method of relating that assurance to data gathered in the development test program when overall system performance is not being measured is also of interest. The methodology described in this report provides a means of quantifying assurance in terms of Bayesian belief probability and relates that belief probability to test data from any source. By capturing the uncertainty in the distribution parameters that comprise key performance parameters when executing the performance simulation, one can create a distribution of belief regarding the system performance parameter based on the body of knowledge about the system at a given time. Through this distribution of belief, one can quantify the belief probability that the true, but unknown, system performance parameter exceeds any given goal.

The example provided in this report shows how the methodology is executed. It is important to note that the example looked only at a subset of the entire problem. When conducting a more comprehensive analysis, one could expect to encounter a variety of distributions and uncertainty regarding any number of parameters that comprise those distributions. There is still much work to be completed before this methodology is a viable tool for use in Army missile system development programs. The attractive feature of this work is its potential for broad application. It is not limited to assessing belief probability in P_k, but rather is applicable to any simulation which utilizes stochastic inputs and processes to develop output.

FUTURE WORK

The focus of future efforts will be to fully develop the methodology to include a variety of distributions (i.e., exponential, log-normal, etc.) and combinations of uncertainty in the parameters that comprise those distributions. Although execution of the methodology was manageable for the example conducted in this report, efficient execution will be of paramount importance when utilizing the methodology to support an Army missile system development program. Many of the steps of the methodology for generating the distributions of belief in this report (construction of the histograms, generation of the standard deviation distributions, etc.) were conducted manually. Future efforts will automate these steps with the intent of simplifying the process. It is expected that the biggest obstacle to implementing the methodology is the number of times the system P_k simulation must be executed. One area being pursued as a means of reducing the number of runs required is the use of surface fits to relate the distribution parameter values to P_k instead of making a run for each combination of the distribution parameter values. By implementing these measures for improving efficiency and developing robust methodology for many distributions and combinations of uncertainty, AMSAA feels this tool will be valuable to the Army in assessing the performance achieved by a developmental weapon system when the primary means for quantifying that performance comes from simulation.

There is another source of uncertainty that contributes to the distribution of belief regarding P_k that was not accounted for in this report. As discussed earlier in this report, the P_k simulation is executed in Monte Carlo fashion a fixed number of times (25 replications in this report) to develop an estimate of P_k for a given set of distribution parameters. Because the number of replications is finite, there is always uncertainty as to what the true P_k is even if the uncertainty in the distribution parameters is ignored. Developing methodology to incorporate this element of uncertainty is another focus of future efforts.

Predictive Quality Control Charts

D. H. Olwell*
Department of Mathematical Sciences
United States Military Academy, West Point, NY 10996-1786†

December 10, 1996

Abstract

Standard control charts require substantial historical data to estimate the parameters of the underlying distribution. While that historical data is being accumulated, can one still monitor the process to determine if it is in control? Can we use subsequent data to refine our initial estimates? The question is especially timely if one wishes to apply Statistical Process Control (SPC) methods to short-run processes. In this paper, we present a predictive control scheme for normal variates based on the predictive distribution, $p(y|\mathbf{x})$, which allows continuously improving control charting from the second observation at the latest. We include some novel graphics for SPC. We discuss the advantages of this approach, and give an example.

KEYWORDS: predictive inference, statistical process control, short-run

Introduction

In this paper, we develop methods for statistical process control based on the predictive distribution for normal variates. This allows control methods to be applied almost immediately, instead of waiting for the 20 or 25 rational groups recommended in the literature (Montgomery, 1985). This is particularly advantageous in short run process control, where there may never be extensive historical data. It is also advantageous for long term control, because the predictive distribution continues to be refined as additional data is accumulated. Use of the predictive distribution confers other advantages, which will be discussed.

We note that predictive control schemes were proposed originally for inverse gaussian processes with a non-informative prior distribution by (Olwell, 1996). We extend the idea here to the normal distribution, and include informative prior distributions and some additional graphic measures for the user.

*Olwell is an assistant professor in the Department of Mathematical Sciences, United States Military Academy, West Point, NY 10996-1786. This research was partially supported by the Army Research Laboratory and the Mathematical Sciences Center of Excellence, USMA.

†Approved for public release; distribution unlimited.

Persistent versus sharp change

The methods of this paper focus on detecting an isolated special cause; that is, a one–time sharp departure from the model for the process. Control charts are best for detecting these shifts.

To detect small persistent changes for start-up data, we recommend self-starting CUSUM charts (Hawkins, 1987) or predictive CUSUM charts, currently under development.

In this spirit, we do not develop or advocate supplementary rules for the predictive charts, since CUSUM charts are optimal for detecting persistent model shifts (Moustakides, 1986)

Uncertainty about parameters

Parameters for a controlled process are never known. At best, we may have very precise estimates for them. In this work, we explicitly model the uncertainty about our parameters, and reflect the improved precision in our knowledge of the parameters that comes with more extensive data.

Before we actually begin to collect data about a process, we may have information or beliefs about how the process will behave. These beliefs can be based on similar processes, prototyping and engineering studies, process specification limits, or general belief about how the process "ought" to behave. Very rarely in an industrial setting will we have no idea about the parameters of the distribution before we actually collect data.

We can capture these beliefs by modeling the parameters themselves as random variables. Using a Bayesian approach, we update our beliefs about the parameters as we observe the process.

If we truly have no information about the parameters, or if we wish to be conservative, we can reflect that lack of information by modeling the parameters with a suitably vague prior distribution.

For example, imagine a production process that fills corn flakes boxes. Before the production line is ever operational, we might believe that the true mean weight of the product inserted will be 16.1 ounces, give or take 0.1 ounce. We also might believe that the standard deviation of the process might be 0.5 ounces, give or take 0.25 ounces. Using these opinions, we could model our belief about the unknown mean by saying that $\mu \sim N(16.1, 0.01)$. We could represent our belief about the unknown variance using a Gamma distribution.

If we had no prior information about the behavior of the weight of the product, we could use a very flat prior distribution, letting the variance of our estimates for the mean and standard deviation grow arbitrarily large.

We will discuss a technique for eliciting these prior beliefs.

The key point is that we can and should incorporate these prior opinions into our model for our control scheme.

Predictive distributions

Predictive distributions are based on a Bayesian approach. In this discussion, y will refer to the unknown next observation, while $\mathbf{x}$ will refer to the observation(s) already made. θ will be the unknown process parameter(s).

We model our data using a parametric distribution, $f(\mathbf{x}|\theta)$. For example, we might believe that the observed data follows a normal distribution with unknown parameters. We have a prior opinion about these process parameters, which are not known exactly. That opinion is represented by $p(\theta)$. This opinion can very strong, or it can be vague.

We observe the process, and collect data. This data is used to update our opinions about the parameters, resulting in $p(\theta|\mathbf{x})$. This follows the standard Bayesian approach:

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{\int_\theta p(\mathbf{x}|\theta)p(\theta)d\theta} .$$

We then integrate over the parameter space to obtain a distribution

$$h(y|\mathbf{x}) = \int_\Theta f(\mathbf{x}|\theta)p(\theta|\mathbf{x})d\theta \tag{1}$$

where y is a future observation of the process, and $\mathbf{x}$ is the historical data.

The normal distribution

In this paper, our concern is with processes modeled by the normal distribution.

For ease of computation, we parameterize the normal distribution as $N(\mu, \tau)$, where μ is the mean, and $\tau = 1/\sigma^2$ is the *precision.* This is a standard notation for the normal distribution when applying Bayesian methods.

We will use conjugate priors here to ease the modeling effort. For (μ, τ) we use the Normal – Scaled Chi-Square (NoCh) joint distribution, which has four hyper–parameters. $(\mu, \tau) \sim NoCh(b, c, g, h)$ implies that $\mu|\tau \sim N(b, c\tau)$ and $\tau h \sim \chi^2_g$. Here and subsequently, we follow the notation of Aitchison and Dunsmore [1975]. Note that $c, g,$ and h must all be non-negative.

The roles of b and c are self-evident: to give the center and scale (as a multiple of τ) of the distribution of the mean.

The roles of g and h are a little more obscure. The Scaled Chi-Square distribution is equivalent to a Gamma distribution with parameters $(g/2, h/2)$. It is helpful to remember that, under this prior distribution,

$$E\left(\frac{1}{\tau}\right) = \frac{h}{g-2} \tag{2}$$

$$Var\left(\frac{1}{\tau}\right) = \frac{2h^2}{(g-4)(g-2)^2} \tag{3}$$

This allows us to make statements about the expected value and variance of σ^2 and then, by matching moments, deduce g and h.

For our corn flake example earlier, $E(\sigma) = 0.5$ and $Var(\sigma) = 0.25^2 = 1/16$. We can use Equations 2 and 3 to obtain $g = 12$ and $h = 5$. From this, it follows that $E(\tau) = 2.4$ (and $Var(\tau) = 0.96$). We use $E(\tau)$ to estimate c. In the corn flake example, we had our uncertainty about μ estimated with a standard deviation of 0.1, resulting in a precision of 100. We then solve

$$100 = cE(\tau) = 2.4c$$

resulting in $c = 41.67$. We round down to $c = 40$.

Our final set of hyper-parameters is

$$(b, c, g, h) = (16.1, 40, 12, 5)$$

For those rare situations where we truly have no prior opinion about the parameters, we can use zero values for c, g, and h to reflect our uncertainty.

Given these priors, we still need the posterior distributions for the parameters and the predictive distribution $h(y|\mathbf{x})$. We will use sufficient statistics for the data. For our historical data with k observations, we represent $m = \overline{\mathbf{x}}$, and $v = \sum_{i=1}^{k} (x_i - \overline{\mathbf{x}})^2$. For the future sample, y, of size K we have the sufficient statistics M and V, respectively. Notice we use lower case letters for our observed data, and capital letters for the future unobserved data.

The calculations are extensive, and we will define intermediate terms to simplify the notation. Aitchison and Dunsmore [1975] provide the relevant posterior and predictive distributions and notation:

$$\begin{aligned}
p(\mu, \tau|\mathbf{x}) &\sim NoCh(B, C, G, H) && (4)\\
p(M|\mathbf{x}) &\sim St\left(G, B, \left(\frac{1}{K} + \frac{1}{C}\right)\frac{H}{G}\right) && (5)\\
p(V|\mathbf{x}) &\sim Si(G, K-1, H) && (6)\\
C &= c + k\\
B &= \frac{cb + km}{C}\\
\Delta(c) &= \begin{cases} 0 & (c = 0)\\ 1 & (c > 0)\end{cases}\\
G &= g + K - 1 + \Delta(c)\\
H &= h + v + \frac{ck(m-b)^2}{c+k}
\end{aligned}$$

For $v > 0$, $Si(k, g, h)$ is a Siegel distribution with density

$$f(v; k, g, h) = \frac{v^{(g/2)-1}}{\beta(k/2, g/2)h^{g/2}(1 + v/h)^{(k+g)/2}} \tag{7}$$

$St(k, b, c)$ is a location-scale transformed student distribution, with k degrees of freedom, centered at b, and scaled by the factor c.

We note that the marginal distribution for $\mu|\mathbf{x}$ is a Student's distribution, $St(G, B, H/(CG))$.

While the distributions look formidable, the calculations are easily relegated to a computer. All the user will see in our implemented scheme are 4 charts. Once the prior is established, the operator will only input M, V, and K for each sample.

Calculations are simplified by identities allowing probabilities involving the posterior distributions to be written in terms of incomplete beta functions, as noted in Aitchison and Dunsmore (1975).

The scheme

We elicit a prior distribution for (μ, τ). This requires judgment and process knowledge. If we specify an unnecessarily vague prior distribution, we will be relatively slower to detect out–of–control states until we have accumulated relatively more data. However, if we specify a precise but mis-located prior, we will signal immediately. The advantage to using informative priors is quicker sensitivity to either mis-specified priors or an out-of-control process.

Once the prior is identified, we start the process. For the first sample $\mathbf{x}$, we obtain the posterior distribution for $(\mu, \tau)|\mathbf{x}$ and a predictive distribution for $y|\mathbf{x}$. We then draw the second sample. We find a p–value for the second sample, using the predictive distribution based on the earlier observation(s). If the p–value is too extreme compared to critical p values, we signal an out–of–control situation. Otherwise, we incorporate the second sample into the historical data set, recalculate our historical summary statistics m and v, and construct an updated posterior distribution and predictive distribution.

We obtain our critical p values by asking the decision maker to specify a tolerable average run length (ARL) between false signals. Then in–control, we use symmetric probability limits:

$$P(\text{ false signal}) = \frac{1}{ARL}$$

and

$$p_{\text{lower}} = \frac{1}{2ARL} \text{ and } p_{\text{upper}} = 1 - \frac{1}{2ARL}$$

We continue sampling, checking, incorporating, and recalculating until the process signals. Our results are presented to the decision-maker using charts.

The charts

We maintain four charts. We maintain charts of the marginal distributions $\mu|\mathbf{x}$ and $\tau|\mathbf{x}$. As we gather more data, these should each approach a point distribution. These allow the process manager to see how much uncertainty remains at any point about the parameters of the process.

The third chart is a rescaled plot of the percentiles of M values against the rational group number. To help distinguish extreme M–values, we use the inverse normal transformation of the percentile of each M

value, based on the predictive distribution:

$$Z = \Phi^{-1}(p)$$

In–control, the p–values are uniformly distributed, resulting in $Z \sim N(0, 1)$.

We plot these rescaled percentiles to obtain constant control limits. Without a transformation, we would have to recalculate the control limits for each observation, which is visually distracting and computationally annoying. Finding variable control limits using the predictive distribution is more computationally demanding than finding a percentile, which just involves a simple numerical integration:

$$p = \int_{-\infty}^{M} h(M|\mathbf{x})dM \tag{8}$$

This integral can be expressed as an incomplete beta integral, allowing the use of fast, accurate existing algorithms. This follows from the identity given in Aitchison and Dunsmore,

$$\int_{a}^{\infty} St(k, b, c) = \frac{1}{2} I_{(1+(kc)^{-1}(a-b)^2)^{-1})}(k/2, 1/2)$$

This third chart has nice asymptotic properties. As the size of the historical record increases, $M|\mathbf{x} \rightarrow N(\mu^*, \tau^*/K)$, where μ^* and τ^* are the asymptotic point distributions and K is the rational group size. $\Phi^{-1}F(Y|\mathbf{x})$ asymptotically just studentizes the observations.

The fourth chart is a plot of the rescaled percentiles for V against the rational group number. Asymptotically, $V|X \sim \chi^2_{K-1}$. Similar to the third chart, we plot

$$W = F^{-1}(p)$$

where F is the CDF for the χ^2_{K-1} distribution.

Examples

We use a simulated data set to illustrate the methods for a vague and informed prior.

We then use a data set from Montgomery(1991) to illustrate the method for both an informed and vague prior distribution. The data consists of 25 rational subgroups of size five, measuring the inside diameter for automobile piston rings. The charts behave abnormally, and post-analysis of the entire data set indicates that the data do not follow the assumed distributions.

We have implemented the calculations on QuattroPro for Windows, a commercially available, inexpensive spreadsheet. All of the graphics are imported from the spreadsheet. A copy of the file, which can be used for any data set, is available from the author.

For all of these examples, we have set the ARL at 370, resulting in performance comparable to 3 σ control limits for normal data.

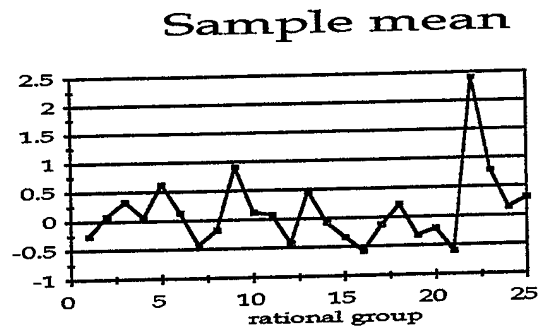


Figure 1: Plot of the sample average for Example 1.

Example 1

We begin with a vague prior: $c, g, h = 0$. Under control, the observations are distributed as $X \sim N(0, 1)$. We draw samples of size 5. We change the distribution only at observation 22 to $N(2, 1)$. The change is noted immediately.

Figure 1 is the plot of sample means. Figure 2 is the plot of sample V. Figure 3 is the plot of the M scores, signaling at observation 22. Figure 4 is the plot of the V scores, which does not signal. Figure 5 is the plot of the distribution of $\mu|\mathbf{X}$ at the end of the data collection. Figure 6 is the plot of the distribution of $\tau|\mathbf{X}$, also after sample 25.

Note from Figures 5 and 6 that even after 25 observations, there is a good deal of uncertainty about the parameters of the process, and the most likely values are not the (here known) true parameter values. This argues for continuing to update these distributions past observation 25, as we would do in the predictive scheme.

Example 2

For this example, we maintain the same model as in Example 1. We change the timing of the model departure to occur earlier in the process, here at observation 6. The process has only 5 samples upon which to base its predictive distribution. Again, we depart to a $N(2, 1)$ distribution.

The departure is again detected immediately. Figure 7 shows the plot of sample averages, Figure 8 the plot of sample V, Figure 9 the plot of M scores, and Figure 10 the plot of V scores.

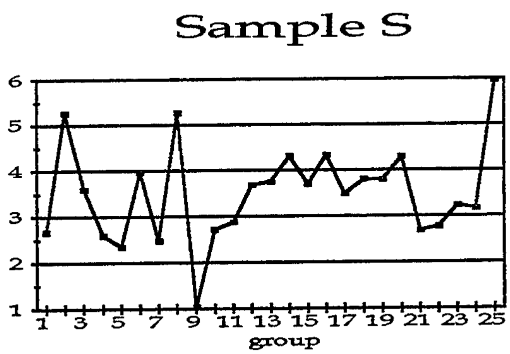


Figure 2: Plot of the sample V for Example 1.

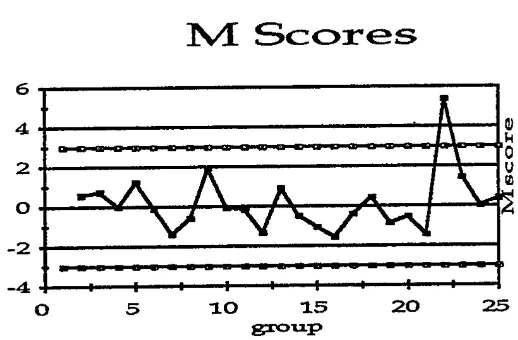


Figure 3: Plot of the M scores for Example 1. Note the signal at observation 22. Also note that there is no M score for the first observation, because we have used a vague prior.

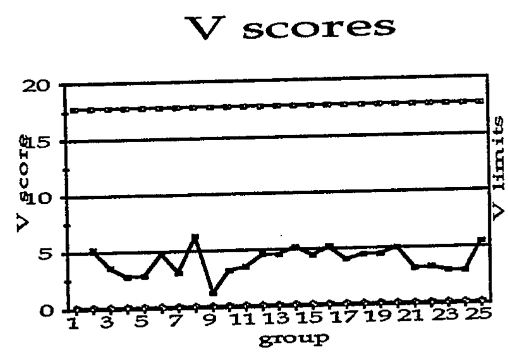


Figure 4: Plot of the V scores for Example 1.

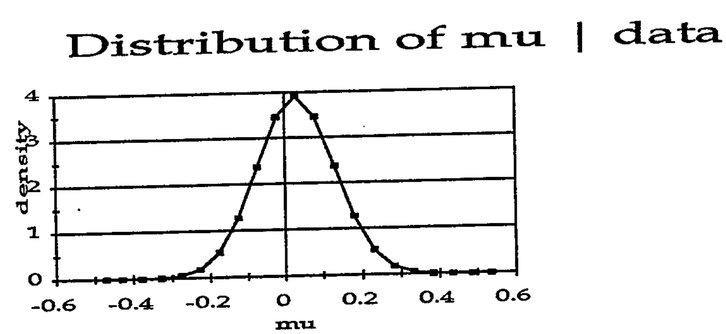


Figure 5: Plot of the posterior distribution for $\mu|\mathbf{X}$ for Example 1, after all 25 observations.

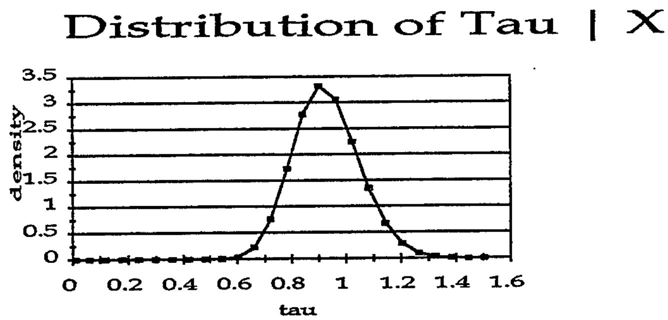


Figure 6: Plot of the posterior distribution for $\tau|\mathbf{X}$ for Example 1, after all 25 observations.

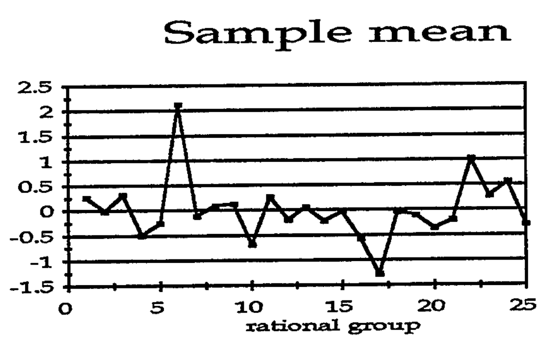


Figure 7: Plot of the sample averages for Example 2.

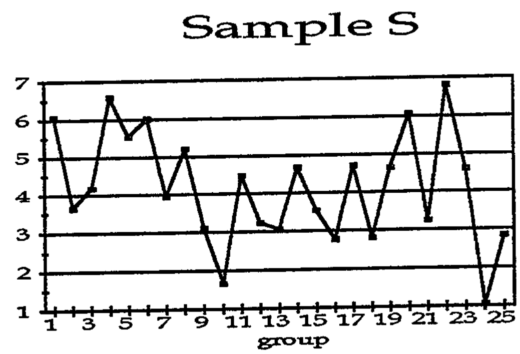


Figure 8: Plot of the sample V for Example 2.

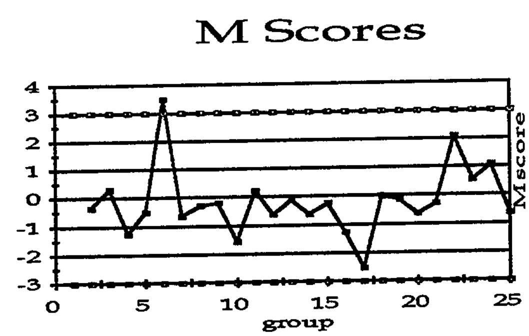


Figure 9: Plot of the M scores for Example 2. Note the signal at observation 6.

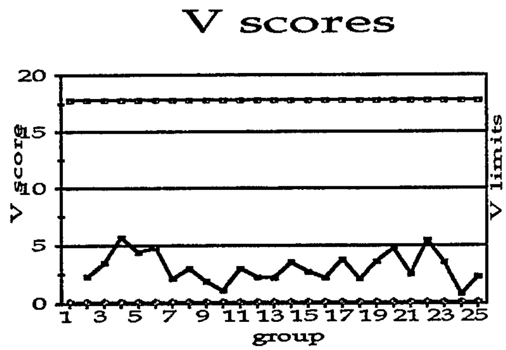


Figure 10: Plot of the V scores for Example 2.

Example 3

We turn to the Montgomery data given in Table 6.3 (Montgomery, 1991), again with a vague prior. Figures 11,12,13,and 14 contain the plots of the sample average, V, M scores, and V scores, respectively.

Note the aberrant behavior of the plot of V scores in Figure 14. This plot shows values apparently much too low. A $q-q$ plot of the sample variances (S^2) is given in Figure 15, and indicates that the sample variances for this published data do not appear to be proportional a χ^2_4 distribution. There are fewer than expected large values of S^2. Accordingly, the plots of the V score do not behave as expected.

While we do not advocate for the use of these charts to detect such model departures, we would not have otherwise been prompted to check the $q-q$ plot for this data.

Again, Figures 16 and 17 indicate how much uncertainty remains about the process mean and precision.

Example 4

This last example shows the effects of strong prior distribution. WE revisit Example 2. We assume $\mu \sim N(0, 100\tau)$. We assume that $E(1/\tau) = 1$ and $Var(1/\tau) = .01$, resulting in $\tau \sim \Gamma(4.02/2, 2.02/2)$. The spreadsheet output for the data is in Figure 18. Note that the plot signals even more strongly for the point which doesn't follow the model. The M scores are separately plotted at Figure 19, for ease of comparison.

Strong prior information improves the sensitivity of the scheme.

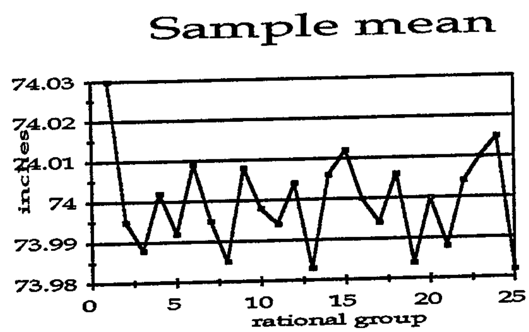


Figure 11: Plot of the sample average for Example 3.

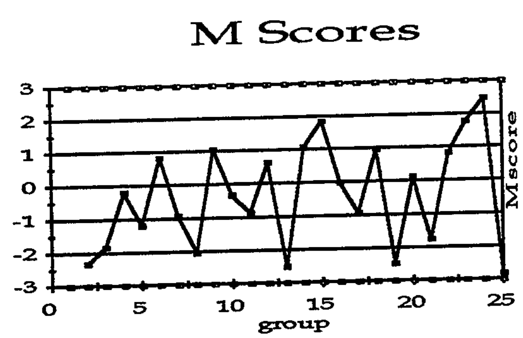


Figure 12: Plot of the sample V for Example 3.

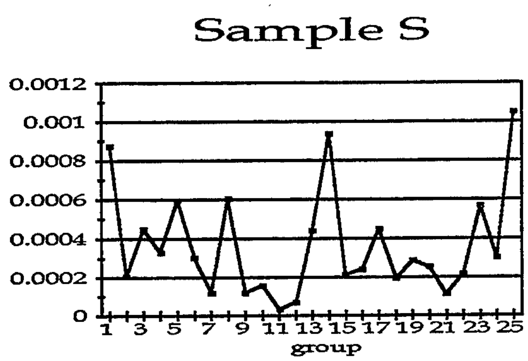


Figure 13: Plot of the M scores for Example 3.

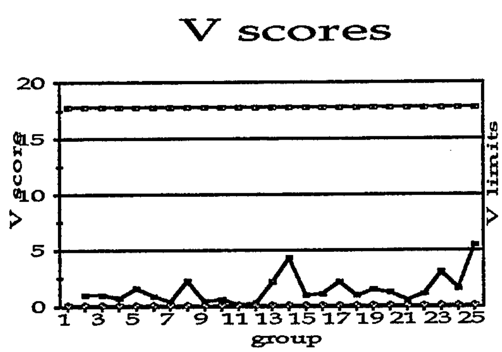


Figure 14: Plot of the V scores for Example 3.

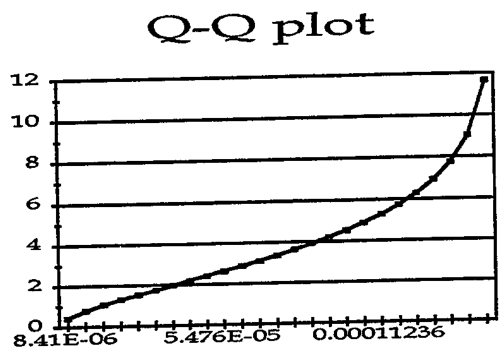


Figure 15: $q - q$ plot of the sample variances against a χ_4^2 for Example 3. Notice the poor fit. The R^2 for the associated regression is 0.85, which is highly significant against the Shapiro-Wilks criteria.

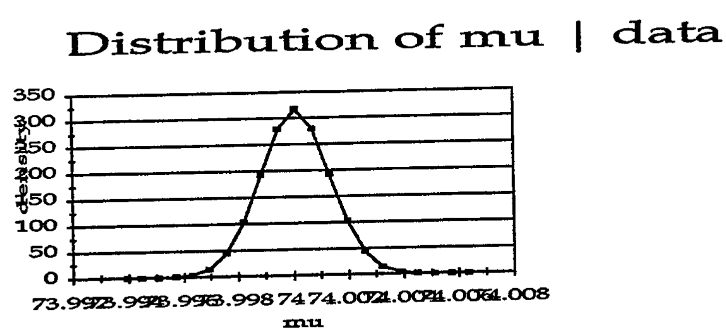


Figure 16: Plot of the posterior distribution for $\mu|\mathbf{X}$ for Example 3, after all 25 observations.

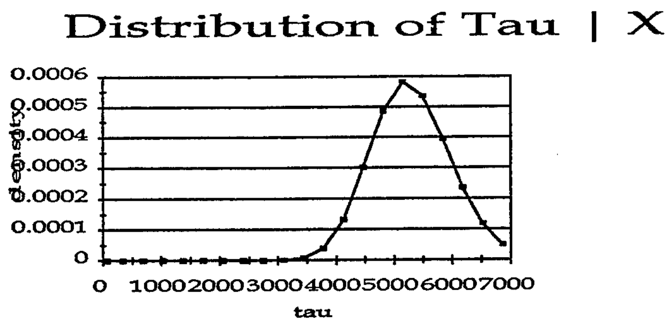


Figure 17: Plot of the posterior distribution for $\tau|\mathbf{X}$ for Example 3, after all 25 observations.

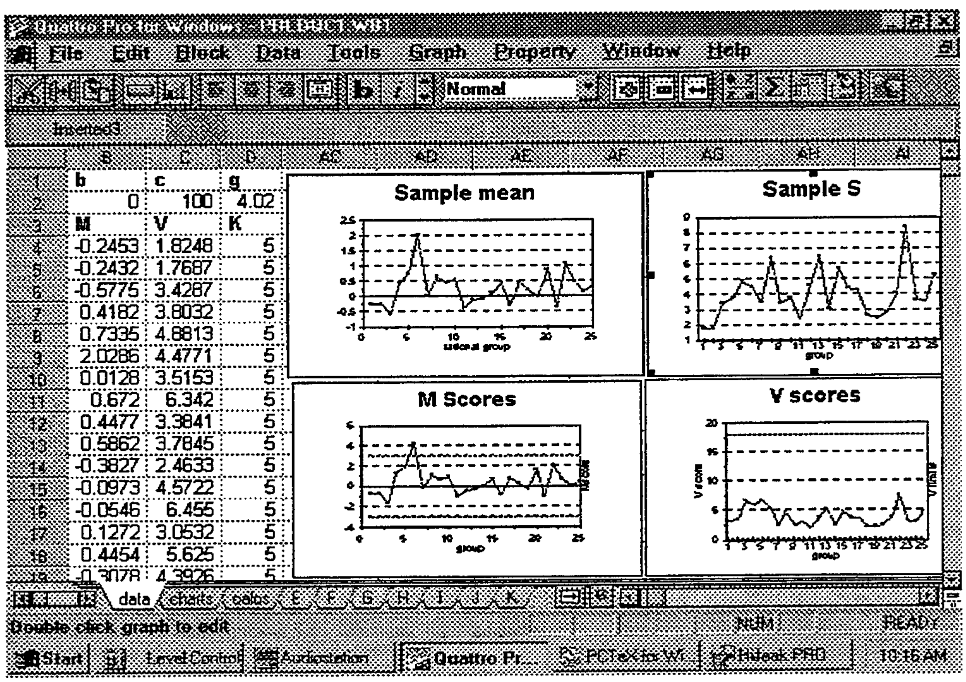


Figure 18: Spreadsheet view of the data from Example Four, with a strong prior. Note the strong signal at observation 6 on the plot of M scores.

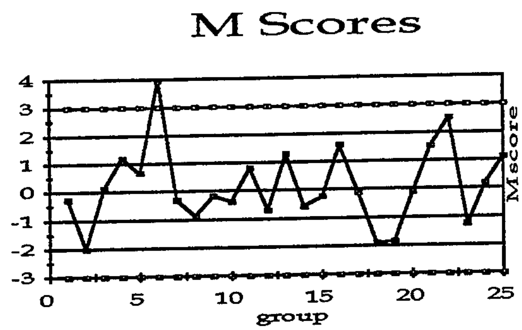


Figure 19: M scores for the data from Example Four, with a strong prior. Note the strong signal at observation 6.

Conclusion

We have introduced a quality control scheme to detect isolated special causes based on the joint predictive distribution for the sample sufficient statistics. This scheme allows us to begin valid SPC immediately, without waiting to accumulate historical data. Additionally, the scheme continues to refine itself as more data is collected, resulting in more precise estimates for the process parameters.

The process is easily implemented on a commercial spreadsheet, as we have done here, and could be added to commercial SPC products with little labor. Once the prior distribution has been estimated, the operator needs only to enter the sample mean, the sample standard deviation or sample V, and the sample size.

The charts implementing the scheme provide useful information about the process behavior and the current sample.

While we have not illustrated this, the charts accept variable sample sizes.

These tools should be adopted by any practitioner confronted with short runs, or a need for continually improving parameter estimates for the process.

References

1. Aitchison, J. and I. R. Dunsmore (1975) *Statistical Prediction Analysis.* Cambridge: Cambridge University Press.

2. Geisser, Seymour (1993) *Predictive Inference: An Introduction* New York: Chapman & Hall.

3. Hawkins, Douglas M. (1987) Self-starting cusums for location and scale. *The Statistician* Vol. 36. pp. 299-315.

4. Montgomery, Douglas C. (1985) *Introduction to Statistical Process Control.* 1st Ed. New York: Wiley.

5. Montgomery, Douglas C. (1991) *Introduction to Statistical Process Control.* 2nd Ed. New York: Wiley.

6. Moustakides, George V. (1986) Optimal Stopping Times for Detecting Changes in Distributions. *Annals of Statistics.* Vol. 14. No. 4. pp. 1379-1387.

7. Olwell, David H. (1996) *Topics in Statistical Process Control.* Ann Arbor: University Microfilms.

Permutation-based, Extrapolated Regression Estimates

(clinical presentation)

David W. Webb

U.S. Army Research Laboratory, APG, MD

Explanation of the Problem

Early in 1996, an electrical engineer in my branch challenged me with data he had collected from a firing test. Twenty-two rounds were fired from a Cannon-Caliber Electromagnetic Gun (CCEMG). Among the variables that he measured from each shot were impact locations (relative to the aimpoint) recorded on yaw cards, and the launch velocity. Launch velocities ranged from 826 m/s to 1,785 m/s; however, when the CCEMG is fully operational its required launch velocity will be 1850 m/s. The task that my co-worker needed assistance with was predicting the dispersion (i.e., the standard deviation of the impacts) of the CCEMG at the full design velocity of 1850 m/s, hereafter denoted σ_{FDV}.

Several challenges confronted this effort. First, there was the problem of deciding how to compute dispersion when there are no exact repeat observations of any velocity. Second, a procedure was needed for extrapolating beyond the observed range of velocities to predict the dispersion at full design velocity.

To address the issue of calculating dispersions in the absence of repeat observations, shots were grouped according to a near-neighbors philosophy; then within each group the impact dispersion and average launch velocity were computed. The engineer had already divided the rounds into four groups. He had designated a low-velocity (800-850 m/s) group consisting of four rounds; two mid-velocity (1,000-1,200 m/s) groups of five and seven rounds; and a high-velocity (1,250-1,800 m/s) group of six rounds. The mid-velocity groups were distinguished by whether or not the bore of the cannon was honed (cleaned) before each firing. Using a stem-and-leaf plot, Table 1 shows the distribution of the 22 launch velocities and the classification of the rounds. [Note: Placing the 1282 m/s round in one of the mid-velocity groups would have made it closer to its neighbors, however it remained in the high-velocity group to stick with the engineer's convention.]

Because there was no prior assumption as to the true physical relationship between velocity and dispersion, a simple linear regression between these variables was used. However, the prospect of using a linear regression of just four data points (one per velocity group) to obtain the prediction of σ_{FDV} seemed quite tenuous. Therefore, rounds within a group were partitioned into smaller subgroups consisting of two rounds each (three rounds for one

of the subgroups if the group size was odd), thereby "creating" more data points for the regression. The number of possible ways to partition the groups into subgroups is shown in Table 2.

Low Velocity	800	26	42	48	48		
	900						
Mid Velocity	1000	15	63	*75*	81	*87*	*87*
	1100	*39*	*40*	76	90	*90*	*90*
	1200	82					
	1300						
High Velocity	1400	49	92				
	1500	15					
	1600	39					
	1700	85					

Table 1: Stem-and-leaf plot of the 22 launch velocities observed in the test. Italicized figures indicate that the bore was honed prior to firing. These rounds form one of the two mid-velocity groups.

By then forming all possible permutatons of the subgroups it would be possible to obtain (3)(10)(105)(15)=47,250 unique regressions of average launch velocity versus dispersion, along with the same number of predictions of σ_{FDV}. Upon ordering these 47,250 estimates, one could then obtain a 90% confidence interval of σ_{FDV}. (Due to the high degree of uncertainty associated with extrapolated estimates, a confidence interval for σ_{FDV} was deemed more appropriate than a point estimate.)

However, at this early stage there was a critical flaw in the analysis. Note in Table 1 that the launch velocities of the four low-velocity rounds are relatively close to each other. Therefore, the use of an average velocity as the dependent variable in a regression, although technically a violation of the usual regression assumptions, should not be of grave concern. On the other hand, the launch velocities of the high-velocity rounds are quite different, ranging from a low of 1282 m/s to a high of 1785 m/s. Is it reasonable to consider rounds with such different launch velocities as near- neighbors and allow their inclusion in the same partition? Probably not. To address this, a closeness criteria was implemented which stated that rounds from within the same group could not be partitioned into the same subgroup if their launch velocities differed by 170 m/s or more. While, admittedly, this value of 170 m/s value may still seem to be too high, it was the smallest difference one could use to still acquire three partitions of two rounds each from the high-velocity group. With this new restriction on the permutations, the

number of possible regressions and estimates of σ_{FDV} dropped from 47,250 to just 1,890 (see Table 3).

Group	Group Size	Subgroup Sizes	Number of Partitions Possible
1 - Low velocity	4	2, 2	$\dfrac{\binom{4}{2}\binom{2}{2}}{2!} = 3$
2A - Mid velocity	5	2, 3	$\binom{5}{2}\binom{3}{3} = 10$
2B - Mid velocity	7	2, 2, 3	$\dfrac{\binom{7}{2}\binom{5}{2}\binom{3}{3}}{2!} = 105$
3 - High velocity	6	2, 2, 2	$\dfrac{\binom{6}{2}\binom{4}{2}\binom{2}{2}}{3!} = 15$

Table 2: Summary of all possible partitions of the four groups into subgroups of size two (and three if necessary).

Group	Group Size	Subgroup Sizes	Number of Partitions Possible
1 - Low velocity	4	2, 2	3
2A - Mid velocity	5	2, 3	6
2B - Mid velocity	7	2, 2, 3	105
3 - High velocity	6	2, 2, 2	15

Table 3: Summary of all possible partitions of the four groups into subgroups of size two (and three, if necessary), when the closeness criteria (no shots within same subgroup having launch velocities differing by 170 m/s or more) is invoked.

A final issue to address was how to use all of the yaw card data in the computation of a dispersion estimate for a particular subgroup. I did not want to discard all data from the nearer yaw cards and use only the most distant yaw cards (where flight perturbations have damped out and the round is most stable). On the other hand, I did not think it wise to give equal weight to impact data from the nearest and the farthest yaw cards, since at close range the flight is not stable and dispersion measurements tend to be inflated.

The formula decided upon as an estimator for the dispersion for a subgroup involved the following: for each yaw card distance, if two or more rounds had impact data, the data was used to compute a dispersion at that particular yaw card distance. Denote this dispersion by s_i, where i indicates yaw card number and i=1,2,...,n. Furthermore let d_i be the distance from the muzzle of the CCEMG to the yaw card station, and n_i be the number of rounds within the subgroup with impact data at yaw card station i. Then the weighted estimate of dispersion for the subgroup is given by the formula,

$$s_w = \sqrt{\frac{\sum d_i(n_i - 1)s_i^2}{\sum d_i(n_i - 1)}}.$$

This "quasi-dispersion" formula is similar to the usual pooling equation for sample standard deviations except that it includes weighting by the distance to each yaw card, so as to minimize the influence of impact data closer to the muzzle.

Table 4 illustrates the use of this formula using data from one of the subgroups of Group 2A. For these rounds, thirteen yaw cards were stationed along the projectile's path to record impact locations. Notice that the first four yaw cards did not yield any impacts (due to improper positioning of the cards) and thus did not contribute to the calculation of s_w.

At this stage of the analysis one proceeds to form all 1,890 partitions of the four groups of data, each time applying linear regression to obtain an estimate of σ_{FDV}. As outlined earlier, after ordering all 1,890 estimates, the outer 5% quantiles are used to form a 90% confidence interval for σ_{FDV}.

		Horizontal Impact Location						
i	d_i	Rnd 1	Rnd 2	Rnd 3	n_i	s_i	$d_i (n_i - 1)$	$d_i (n_i - 1) s_i^2$
1	5.0	n/a	n/a	n/a	0	n/a	n/a	n/a
2	9.8	n/a	n/a	n/a	0	n/a	n/a	n/a
3	15.0	n/a	n/a	n/a	0	n/a	n/a	n/a
4	20.0	n/a	n/a	n/a	0	n/a	n/a	n/a
5	25.0	1.177	2.370	2.535	3	0.741	50.0	27.46
6	29.9	0.968	2.120	2.857	3	0.952	59.8	54.20
7	170.1	n/a	n/a	3.285	1	n/a	n/a	n/a
8	175.0	n/a	n/a	3.388	1	n/a	n/a	n/a
9	180.0	1.892	2.766	n/a	2	0.618	180.0	68.75
10	185.0	1.800	2.350	3.314	3	0.766	370.0	217.31
11	190.3	1.832	2.427	3.447	3	0.817	380.6	253.90
12	220.8	n/a	n/a	3.374	1	n/a	n/a	n/a
13	222.0	1.946	2.631	3.545	3	0.802	444.0	285.75
							$\Sigma = 1484.4$	$\Sigma = 907.37$
							$s_w = \sqrt{907.37} \div \sqrt{1484.4} = 0.782$	

Table 5: Sample calculation of "quasi-dispersion"

<u>Questions for the panel:</u>

1. Is the interval formed truly a confidence interval for the dispersion at full design velocity, or is it more akin to a prediction interval for a single observation, or is it something else?

2. The decision to use simple linear regression was made to keep the analysis as uncomplicated as possible, given the errors in the dependent variable. Is this a reasonable choice, despite the fact that physics might suggest using either transformed variables or a more complex regression model?

3. Is the all-possible-permutations approach to resampling the data adequate, or should a bootstrap method have been used to randomly resample?

4. Is the use of distance as weighting factor in my "quasi dispersion" ill-advised?

5. Are there other strategies for estimating σ_{FDV} to recommend?

INTENTIONALLY LEFT BLANK.

POWER STUDY BASED ON SIMULATIONS USING THE WILCOXON SIGNED-RANK TEST

Thomas R. Walker
US Army Aberdeen Test Center
ATTN: STEAC-EN-AA
Aberdeen Proving Ground, MD 21005-5059
email: twalker@atc.army.mil
410-278-7543
DSN 298-7543

ABSTRACT

The Wilcoxon Signed-Rank Test is a nonparametric test for the equivalence of population medians whose statistic is based on the differences between observations in an ordered pair. This test could be used to determine if the skill level of a soldier before and after training is significantly different. The Wilcoxon Signed-Rank Test is well documented in numerous statistics books such as Conover's *Practical Nonparametric Statistics* and others.

The author has performed a limited simulation study and seeks panel comment.

OBJECTIVE

The objective of this work is to determine what is "lost" (less powerful) statistically as the sample size decreases when using the Wilcoxon Signed-Rank Test (i.e. How much do you "lose" statistically if the number of soldiers (comparisons) available for a given test decreases 20 to 12, 20 to 18, 30 to 10, and so on?)

BACKGROUND INFORMATION

A simulation was done to compare the results of the Wilcoxon Signed-Rank Test when the sample sizes (number of soldiers) changed. The sample sizes used in this simulation were 10 to 96 in increments of 2. The probability of detecting a difference was calculated for the sample sizes over various given probabilities (.5 to .9). (For example, a given probability of .700 implies for sample sizes of 20, 18, and 12 that the average number of positive differences is approximately 14 (.7 * 20), 12.6 (.7 * 18), and 8.4 (.7 * 12), respectively. In other words, approximately 14 of the 20 measurements in the first group are greater than the measurements in the second group. Also, the number of differences would have to be integers, but for comparisons of different sample sizes, the percent of the various sample sizes was used.) The simulation was done by the following method:

(1) N uniform random numbers (0 to 1) were generated (N = 10 to 96 in increments of 2).

(2) The integers from 1 to N were put in a column next to the random numbers.

(3) The various "given probabilities" (.5, .6, .7, .8, and .9) were compared with the random numbers. If the random number is less than the "given probability", the comparison is different, otherwise the comparison is the same.

(4) If the comparisons were different, the integers (from 1 to N) were considered negative, otherwise positive.

(5) The quotient of the "sum of the integers" and the "square root of the sum of the squares of the integers" was determined.

(6) This procedure was done 1000 times.

(7) A count was done to determine the "number of the quotients" greater than 1.645 (z value of upper 0.95 level) or less than -1.645 (z value of lower .95 level). This count was divided by the number of simulations (1000). This quotient is the probability that the two samples are different at the .10 significance level (alpha = .10) for a given probability.

(8) This whole procedure was repeated 50 times.

The averages of the 50 quotients for the various "given probabilities" (.5 to .9) are shown in Table 1 and Figure 1. The minimum, maximum, and average of these quotients are shown in Figures 2 through 6 and Appendix A.

TABLE 1. PROBABILITY OF DETECTING THAT ITEMS ARE DIFFERENT (APLPH = .10) FOR SAMPLE SIZES OF N FOR GIVEN PROBABILITIES OF .5, .6, .7, .8 AND .9

	Given Probability				
N	0.5	0.6	0.7	0.8	0.9
10	0.106	0.161	0.320	0.573	0.851
12	0.114	0.175	0.363	0.636	0.897
14	0.105	0.178	0.390	0.687	0.933
16	0.102	0.192	0.429	0.736	0.958
18	0.099	0.189	0.445	0.770	0.972
20	0.096	0.196	**0.477**	0.807	0.984
22	0.098	0.208	0.508	0.841	0.990
24	0.100	0.221	0.544	0.873	0.994
26	0.100	0.228	0.570	0.894	0.996
28	0.098	0.236	0.597	0.912	0.998
30	0.101	0.249	0.625	0.928	0.999
32	0.103	0.257	0.648	0.940	1.000
34	0.103	0.270	0.673	0.950	0.999
36	0.101	0.277	0.696	0.960	1.000
38	0.103	0.289	0.717	0.968	1.000
40	0.102	0.297	0.736	0.974	1.000
42	0.099	0.303	0.751	0.978	1.000
44	0.101	0.315	0.775	0.983	1.000
46	0.099	0.319	0.785	0.986	1.000
48	0.096	0.336	0.805	0.989	1.000
50	0.101	0.341	0.817	0.991	1.000
52	0.101	0.349	0.829	0.993	1.000
54	0.101	0.358	0.841	0.994	1.000
56	0.099	0.368	0.854	0.996	1.000
58	0.102	0.378	0.864	0.996	1.000
60	0.100	0.384	0.873	0.997	1.000
62	0.101	0.392	0.881	0.997	1.000
64	0.094	0.409	0.900	0.997	1.000
66	0.099	0.408	0.898	0.998	1.000
68	0.099	0.418	0.907	0.998	1.000
70	0.100	0.430	0.915	0.999	1.000
72	0.100	0.434	0.921	0.999	1.000
74	0.101	0.445	0.928	0.999	1.000
76	0.101	0.453	0.933	0.999	1.000
78	0.102	0.462	0.939	0.999	1.000
80	0.096	0.475	0.943	0.999+	1.000
82	0.101	0.478	0.947	0.999+	1.000
84	0.104	0.486	0.951	0.999+	1.000
86	0.102	0.493	0.955	0.999+	1.000
88	0.101	0.503	0.959	0.999+	1.000
90	0.102	0.508	0.962	0.999+	1.000
92	0.103	0.516	0.966	0.999+	1.000
94	0.103	0.523	0.969	0.999+	1.000
96	0.094	0.530	0.972	0.999+	1.000

Notes: For a sample size of 20 with a given probability of .7, the probability of detecting that the items are different is .477 at the .10 significance level.

Figure 1. Probability of detecting item differences for various "given probabilities" and sample sizes

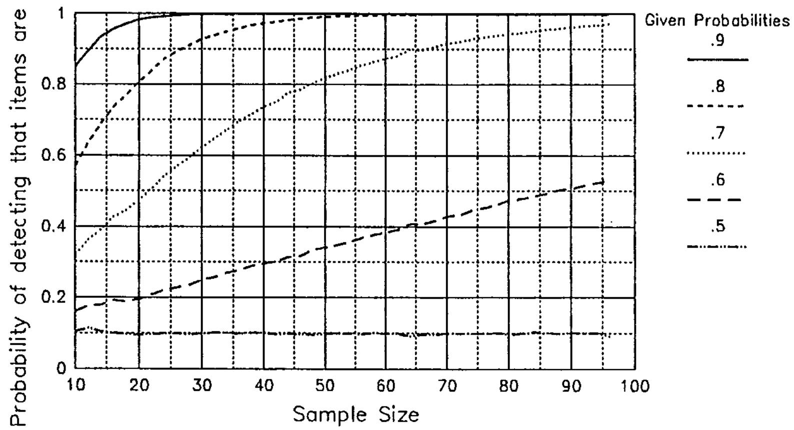


Figure 2. Probability of detecting that the items are different when the given probability is 0.50 for sample sizes between 10 and 96

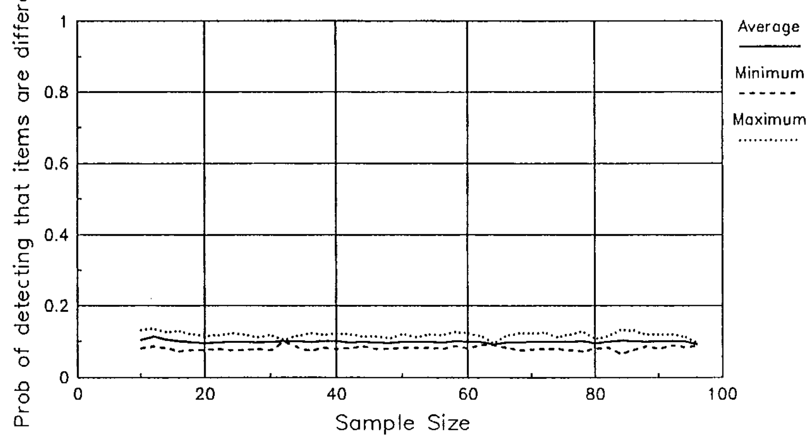


Figure 3. Probability of detecting that the items are different when the given probability is 0.60 for sample sizes between 10 and 96

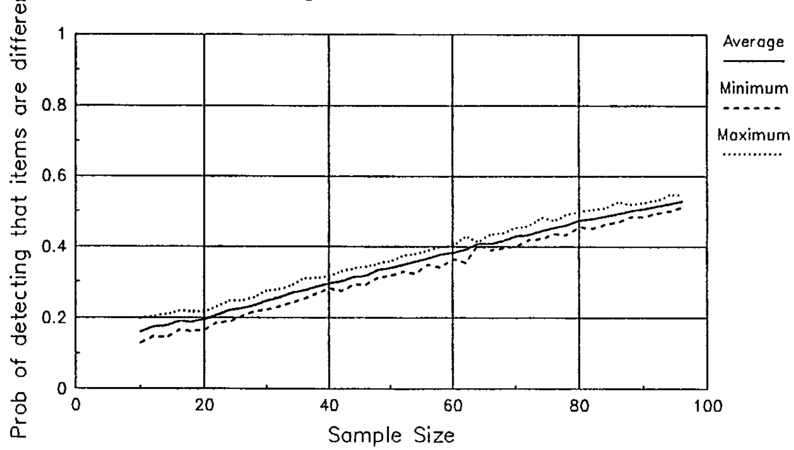


Figure 4. Probability of detecting that the items are different when the given probability is 0.70 for sample sizes between 10 and 96

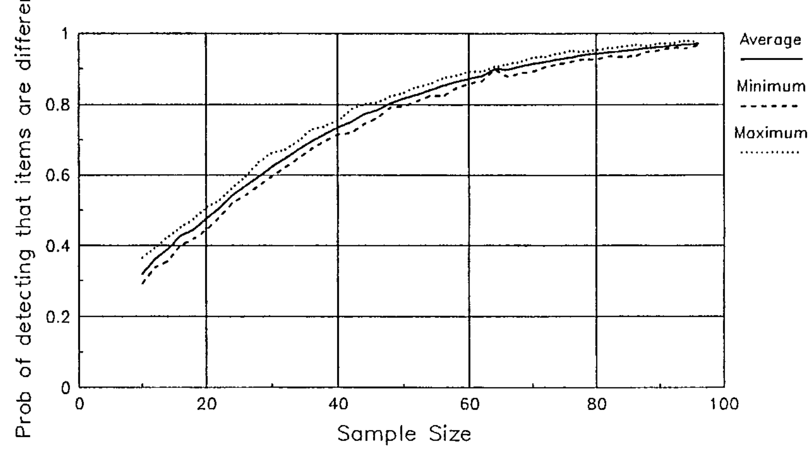


Figure 5. Probability of detecting that the items are different when the given probability is 0.80 for sample sizes between 10 and 96

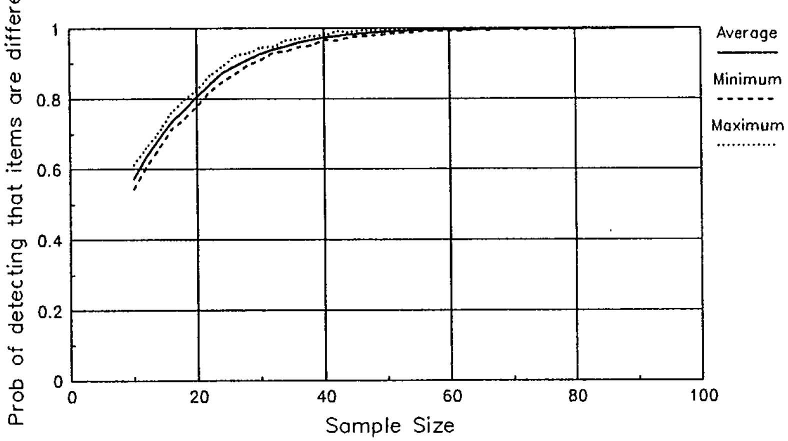
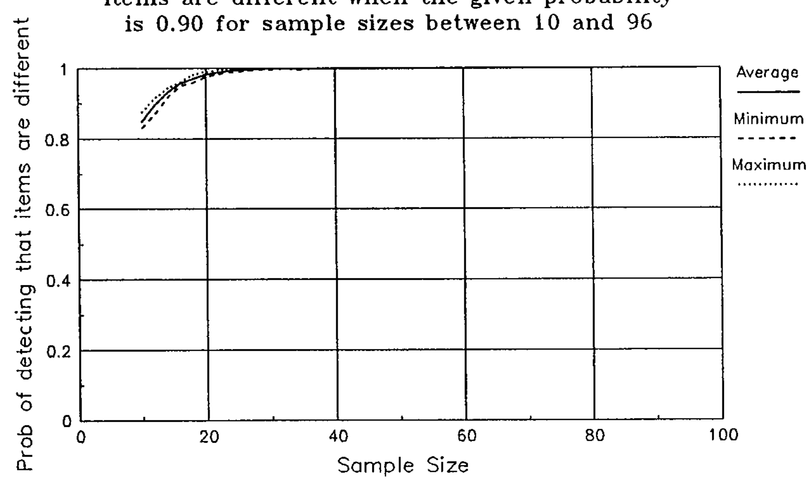


Figure 6. Probability of detecting that the items are different when the given probability is 0.90 for sample sizes between 10 and 96



RESULTS/CONCLUSIONS OF SIMULATIONS

(1) For "given probabilities" of x and 1 - x, the probability of finding that the items are different have similar results due to symmetry (i.e. If the given probability is .70, the probability that the items are different would give the same results as when the given probability is .30 (1 - .70) with a large enough sample size.) Therefore, all simulations were done with "given probabilities" greater than or equal to .5.

(2) For all "given probabilities" greater than .5, the probability of detecting that the items are different will approach 1 as the sample size increases. However, for a "given probability" of exactly .5, the probability of detecting that the items are different approaches alpha of .10.

(3) As the "given probability" increases from .5 the probability that the items are different increases and the sample size required to show a difference decreases (i.e. For a "given probability" = .7, N = 30, the probability that the items are different is .625. While for a "given probability" = .8, N = 12, the probability that the items are different is .636.)

(4) For a "given probability" of .9 and a small sample size of 10, the probability of detecting that the items are different is at least .85. With sample sizes greater than 30, the probability of detecting that the items are different is greater than .999.

(5) For a "given probability" of .6 and a small sample size of 10, the probability of detecting that the items are different is less than .2. In order to detect that the items are different with a probability of at least .50, the sample size would have to be approximately 88.

(6) When the sample size (number of soldiers, etc) decreases from 20 to 12, 20 to 18, 30 to 10 for a "given probability" of .7, the probabilities of detecting that the items are different decreases 24% (.477 to .363), 7% (.477 to .445), and 49% (.625 to .320), respectively.

In short, very little is "lost" statistically when the sample size decreases from 20 to 18 but not so for 20 to 12 and 30 to 10.

ACKNOWLEDGEMENTS

I wish to thank V. V. Visnaw, W. J. Conover, and Huan Le for providing technical comments/input/simulation results.

Appendix A

	Given Probability = .5				Given Probability = .6				Given Probability = .7			
N	Avg	Std	Min	Max	Avg	Std	Min	Max	Avg	Std	Min	Max
10	0.106	0.0122	0.082	0.134	0.161	0.0123	0.129	0.198	0.320	0.0146	0.294	0.365
12	0.114	0.0127	0.088	0.137	0.175	0.0135	0.147	0.206	0.363	0.0148	0.338	0.395
14	0.105	0.0128	0.082	0.128	0.178	0.0146	0.145	0.211	0.390	0.0166	0.357	0.425
16	0.102	0.0148	0.073	0.130	0.192	0.0147	0.166	0.220	0.429	0.0130	0.401	0.452
18	0.099	0.0094	0.077	0.121	0.189	0.0140	0.162	0.218	0.445	0.0133	0.420	0.475
20	0.096	0.0088	0.078	0.117	0.196	0.0118	0.166	0.218	0.477	0.0158	0.445	0.510
22	0.098	0.0094	0.080	0.118	0.208	0.0109	0.188	0.233	0.508	0.0129	0.481	0.529
24	0.100	0.0109	0.078	0.124	0.221	0.0134	0.189	0.248	0.544	0.0125	0.523	0.565
26	0.100	0.0092	0.078	0.120	0.228	0.0109	0.206	0.248	0.570	0.0129	0.543	0.597
28	0.098	0.0086	0.080	0.113	0.236	0.0095	0.218	0.257	0.597	0.0144	0.572	0.638
30	0.101	0.0068	0.078	0.119	0.249	0.0114	0.223	0.278	0.625	0.0143	0.598	0.664
32	0.103	0.0015	0.100	0.106	0.257	0.0116	0.233	0.281	0.648	0.0138	0.623	0.673
34	0.103	0.0067	0.087	0.116	0.270	0.0116	0.243	0.291	0.673	0.0122	0.648	0.699
36	0.101	0.0112	0.076	0.124	0.277	0.0132	0.254	0.312	0.696	0.0121	0.675	0.732
38	0.103	0.0088	0.085	0.120	0.289	0.0120	0.268	0.313	0.717	0.0130	0.696	0.740
40	0.102	0.0120	0.080	0.124	0.297	0.0094	0.283	0.317	0.736	0.0086	0.717	0.755
42	0.099	0.0100	0.082	0.121	0.303	0.0109	0.276	0.332	0.751	0.0144	0.721	0.787
44	0.101	0.0066	0.088	0.115	0.315	0.0099	0.293	0.338	0.775	0.0152	0.746	0.803
46	0.099	0.0086	0.080	0.115	0.319	0.0101	0.294	0.344	0.785	0.0103	0.763	0.806
48	0.096	0.0067	0.081	0.110	0.336	0.0089	0.314	0.354	0.805	0.0065	0.793	0.822
50	0.101	0.0086	0.085	0.122	0.341	0.0093	0.319	0.361	0.817	0.0089	0.796	0.834
52	0.101	0.0079	0.085	0.113	0.349	0.0110	0.329	0.375	0.829	0.0081	0.811	0.849
54	0.101	0.0095	0.085	0.120	0.358	0.0127	0.325	0.380	0.841	0.0079	0.824	0.858
56	0.099	0.0096	0.080	0.118	0.368	0.0114	0.349	0.389	0.854	0.0120	0.825	0.874
58	0.102	0.0102	0.088	0.128	0.378	0.0138	0.342	0.399	0.864	0.0087	0.846	0.882
60	0.100	0.0104	0.082	0.123	0.384	0.0101	0.365	0.401	0.873	0.0087	0.859	0.893
62	0.101	0.0053	0.091	0.114	0.392	0.0168	0.355	0.430	0.881	0.0064	0.867	0.894
64	0.094	0.0013	0.091	0.096	0.409	0.0022	0.405	0.414	0.900	0.0016	0.898	0.905
66	0.099	0.0072	0.085	0.116	0.408	0.0104	0.390	0.435	0.898	0.0082	0.878	0.913
68	0.099	0.0118	0.078	0.124	0.418	0.0099	0.397	0.439	0.907	0.0070	0.887	0.920
70	0.100	0.0090	0.079	0.124	0.430	0.0097	0.401	0.453	0.915	0.0080	0.892	0.932
72	0.100	0.0145	0.081	0.126	0.434	0.0096	0.420	0.460	0.921	0.0062	0.906	0.935
74	0.101	0.0071	0.081	0.113	0.445	0.0132	0.424	0.482	0.928	0.0073	0.913	0.943
76	0.101	0.0100	0.078	0.120	0.453	0.0095	0.435	0.473	0.933	0.0067	0.921	0.950
78	0.102	0.0146	0.073	0.130	0.462	0.0116	0.434	0.491	0.939	0.0051	0.928	0.948
80	0.096	0.0060	0.082	0.108	0.475	0.0126	0.455	0.501	0.943	0.0077	0.928	0.955
82	0.101	0.0080	0.085	0.116	0.478	0.0144	0.452	0.507	0.947	0.0054	0.936	0.959
84	0.104	0.0180	0.067	0.132	0.486	0.0108	0.463	0.510	0.951	0.0061	0.935	0.963
86	0.102	0.0145	0.077	0.131	0.493	0.0126	0.469	0.528	0.955	0.0063	0.937	0.968
88	0.101	0.0072	0.088	0.121	0.503	0.0088	0.485	0.520	0.959	0.0036	0.948	0.965
90	0.102	0.0085	0.080	0.121	0.508	0.0085	0.484	0.525	0.962	0.0046	0.952	0.972
92	0.103	0.0084	0.090	0.120	0.516	0.0096	0.496	0.531	0.966	0.0032	0.959	0.972
94	0.103	0.0063	0.087	0.115	0.523	0.0107	0.501	0.548	0.969	0.0058	0.958	0.982
96	0.094	0.0016	0.088	0.096	0.530	0.0108	0.513	0.548	0.972	0.0024	0.967	0.977

	Given Probability = .8				Given Probability = .9			
N	Avg	Std	Min	Max	Avg	Std	Min	Max
10	0.573	0.0149	0.544	0.612	0.851	0.0096	0.832	0.877
12	0.636	0.0108	0.608	0.661	0.897	0.0095	0.871	0.916
14	0.687	0.0121	0.660	0.706	0.933	0.0062	0.919	0.945
16	0.736	0.0121	0.714	0.763	0.958	0.0025	0.952	0.963
18	0.770	0.0126	0.744	0.799	0.972	0.0047	0.960	0.982
20	0.807	0.0088	0.781	0.825	0.984	0.0028	0.977	0.991
22	0.841	0.0107	0.820	0.866	0.990	0.0027	0.985	0.996
24	0.873	0.0093	0.849	0.891	0.994	0.0020	0.989	0.997
26	0.894	0.0119	0.870	0.923	0.996	0.0020	0.992	1.000
28	0.912	0.0089	0.896	0.930	0.998	0.0010	0.996	1.000
30	0.928	0.0080	0.911	0.946	0.999	0.0010	0.997	1.000
32	0.940	0.0058	0.931	0.948	1.000	0.0005	0.999	1.000
34	0.950	0.0074	0.935	0.966	0.999	0.0009	0.997	1.000
36	0.960	0.0058	0.948	0.971	1.000	0.0006	0.998	1.000
38	0.968	0.0055	0.952	0.978	1.000	0.0004	0.999	1.000
40	0.974	0.0037	0.967	0.981	1.000	0.0000	1.000	1.000
42	0.978	0.0047	0.966	0.992	1.000	0.0003	0.999	1.000
44	0.983	0.0036	0.975	0.990	1.000	0.0002	0.999	1.000
46	0.986	0.0031	0.979	0.994	1.000	0.0000	1.000	1.000
48	0.989	0.0037	0.981	0.994	1.000	0.0000	1.000	1.000
50	0.991	0.0025	0.984	0.996	1.000	0.0000	1.000	1.000
52	0.993	0.0022	0.988	0.996	1.000	0.0000	1.000	1.000
54	0.994	0.0025	0.989	0.999	1.000	0.0000	1.000	1.000
56	0.996	0.0015	0.993	0.999	1.000	0.0000	1.000	1.000
58	0.996	0.0015	0.992	0.999	1.000	0.0000	1.000	1.000
60	0.997	0.0017	0.993	1.000	1.000	0.0000	1.000	1.000
62	0.997	0.0019	0.993	1.000	1.000	0.0000	1.000	1.000
64	0.997	0.0002	0.997	0.998	1.000	0.0000	1.000	1.000
66	0.998	0.0013	0.995	1.000	1.000	0.0000	1.000	1.000
68	0.998	0.0011	0.996	1.000	1.000	0.0000	1.000	1.000
70	0.999	0.0011	0.996	1.000	1.000	0.0000	1.000	1.000
72	0.999	0.0008	0.997	1.000	1.000	0.0000	1.000	1.000
74	0.999	0.0006	0.998	1.000	1.000	0.0000	1.000	1.000
76	0.99896	0.0010	0.997	1.000	1.000	0.0000	1.000	1.000
78	0.99930	0.0008	0.998	1.000	1.000	0.0000	1.000	1.000
80	0.99952	0.0006	0.998	1.000	1.000	0.0000	1.000	1.000
82	0.99954	0.0005	0.999	1.000	1.000	0.0000	1.000	1.000
84	0.99962	0.0005	0.998	1.000	1.000	0.0000	1.000	1.000
86	0.99966	0.0006	0.998	1.000	1.000	0.0000	1.000	1.000
88	0.99976	0.0004	0.999	1.000	1.000	0.0000	1.000	1.000
90	0.99970	0.0005	0.999	1.000	1.000	0.0000	1.000	1.000
92	0.99976	0.0004	0.999	1.000	1.000	0.0000	1.000	1.000
94	0.99976	0.0004	0.999	1.000	1.000	0.0000	1.000	1.000
96	0.99952	0.0005	0.999	1.000	1.000	0.0000	1.000	1.000

RESULTS/CONCLUSIONS OF SIMULATIONS

(1) For "given probabilities" of x and 1 - x, the probability of finding that the items are different have similar results due to symmetry (i.e. If the given probability is .70, the probability that the items are different would give the same results as when the given probability is .30 (1 - .70) with a large enough sample size.) Therefore, all simulations were done with "given probabilities" greater than or equal to .5.

(2) For all "given probabilities" greater than .5, the probability of detecting that the items are different will approach 1 as the sample size increases. However, for a "given probability" of exactly .5, the probability of detecting that the items are different approaches alpha of .10.

(3) As the "given probability" increases from .5 the probability that the items are different increases and the sample size required to show a difference decreases (i.e. For a "given probability" = .7, N = 30, the probability that the items are different is .625. While for a "given probability" = .8, N = 12, the probability that the items are different is .636.)

(4) For a "given probability" of .9 and a small sample size of 10, the probability of detecting that the items are different is at least .85. With sample sizes greater than 30, the probability of detecting that the items are different is greater than .999.

(5) For a "given probability" of .6 and a small sample size of 10, the probability of detecting that the items are different is less than .2. In order to detect that the items are different with a probability of at least .50, the sample size would have to be approximately 88.

(6) When the sample size (number of soldiers, etc) decreases from 20 to 12, 20 to 18, 30 to 10 for a "given probability" of .7, the probabilities of detecting that the items are different decreases 24% (.477 to .363), 7% (.477 to .445), and 49% (.625 to .320), respectively.

In short, very little is "lost" statistically when the sample size decreases from 20 to 18 but not so for 20 to 12 and 30 to 10.

ACKNOWLEDGEMENTS

I wish to thank V. V. Visnaw, W. J. Conover, and Huan Le for providing technical comments/input/simulation results.

INTENTIONALLY LEFT BLANK.

IMPROVING USE OF STATISTICS IN ARMY TEST AND EVALUATION

Herman Chernoff
Harvard University
Cambridge, MA 02138

ABSTRACT

Three topics are discussed. First, there is a need for more well trained statisticians at the Department of Defense. Second is an outline of an approach to deal with the problem of operational testing of a system for potential use in many environments, when tests can be carried out in very few, one, two or three environments. This approach suggests the use of multidimensional scale analysis as one of the tools with which to reduce the scope of the problem. Finally the third topic is "How large should the sample size be?" It is shown how small sample sizes make it difficult to demonstrate reliability with confidence. For testing hypotheses where sample sizes have to be decided nonsequentially, in advance of experimentation, a Bayesian approach is helpful, and the use of normal theory approximations allow one to use some insightful graphs for approximate solutions.

INTRODUCTION

I was somewhat surprised to discover the title of my lecture, since it suggests a more global view than I am accustomed to taking. My preference is to concentrate on a rather narrow topic and hope that the discussion of that will suggest wider applications. In view of the title, let me address three rather separate subjects. These are the role of statisticians in defense, a special problem in operational testing, and a problem of hypothesis testing.

Since I will be preaching to the converted on the first topic, I will keep that brief. The operational testing problem, sometimes called "Dubin's challenge" is that of selecting a few (two or three) testing environments in which to test a system which is potentially required to function in many environments. The third topic involves a couple of examples which address the question, "How large should the sample size be?", a question which comes up frequently in testing.

The last two topics might be more properly entitled "Is there a free lunch?" I suspect that some administrators responsible for allocating funds for testing may find some of the conclusions disturbing.

Lest I be mistaken for more of an expert than I am on this topic, let me describe my background. I am a Professor of Statistics who worked on an applied ONR contract for many years during which I had contact with a variety of applications of statistics in defense work. I now serve as a member of an NRC (National Research Council) panel on Operational Testing which has been studying and hopes soon to report on several issues in Operational Testing. From these experiences I have some exposure to the real problems in Army Test and Evaluation but that exposure lacks the depths that come from the healthy experiences of being forced to deal with specific examples from beginning to end.

STATISTICIANS IN DEFENSE

A defense department operational test of a system is typically expensive and involves the use of talented physicists and engineers to devise and install appropriate sensors. It is a common saying among physicists that if an experiment needs statistical analysis, it is the wrong experiment. That saying is based on a historical luxury in science, where the major cost of experimentation was that of setting up the experiment. After the experiment is set up, it is relatively costless to replicate it as many times as necessary to get a desired level of accuracy. That luxury is no longer as available as it used to be, and it certainly is not available in operational testing. But a consequence of the attitude described above is that most scientists are statistically naive and unaffected by most of the twentieth century revolutions in statistical theory and practice. The advances in experimental design, sequential analysis and decision theory, among many others, are not appreciated by many of the decision makers in operational testing.

If we examine the types of statistical issues that arise and the personnel available to deal with these problem, there seems to be a mismatch. Rather few of the people who are responsible for facing such issues have more than a trivial background in statistics. Under proper guidance, they can be trained to deal with a variety of standard problems. However issues of experimental design abound, and there are very few people with enough talent to absorb the results of a three day workshop on that topic and apply them creatively. Some healthy and sustained exposure to the theory and practice of statistics is almost always necessary to be successful.

Finally the real world involves unusual and unexpected variations of standard problems. To deal with these problems requires the training and talent to be able to recognize which rules could and should be broken and how to adapt.

In summary the defense department would profit from employing more well trained and capable statisticians. Statistical laymen with the benefit of a handbook or two and a couple of three day workshops will rarely be able to do the job without experienced backup.

EXPERIMENTAL DESIGN IN OPERATIONAL TESTING UNDER LIMITED EXPERIMENTATION

INTRODUCTION

How should one treat the problem of testing a type of equipment in the field when the equipment is expected to be used in several of a large variety of potential environments and funds are only available to test under very few environments? In the following I describe an approach to this problem which, unfortunately, fails to deal with one of the major functions of operational testing. That function is that of discovering the surprises that quickly locate unanticipated but glaring weaknesses, the removal of which makes for an improved product. The approach is described through an example. While the example is artificial, I believe that it is sufficiently realistic to permit discussion of the important ideas. After presenting the "results," I will review the various steps to indicate issues and alternatives. In the end this presentation can serve as a basis for soliciting a slightly more realistic problem on which the issues can be examined with more care.

Finally, this problem has ramifications in a wide range of applications. For example, in testing software, one may subject the software to thousands of test scenarios. Nevertheless, the

set of possible applications is enormously larger than what we may be able to apply in a test with limited time.

THE EXAMPLE

The example is an electric generator which may be required to function in many environments. We shall list 8 possible environments and evaluate these by using numerical values between 0 and 10 for each of 18 stress variables. High values indicate large perceived stress. Thus our stress matrix is an 18*8 matrix $A = ||a_{ij}||$ listed in Table 1 with a description of the rows (variables) and columns (environments) in Table 2. Using the measure of distance or dissimilarity where $d_{jj'}$ is the distance between the j and j', columns, i.e.

$$d_{jj'} = \left\{ \sum_{i=1}^{18} (a_{ij} - a_{ij'})^2 \right\}^{1/2}$$

we have $D = ||d_{jj'}||$ in Table 3.

The Splus routine cmdscale($D, k = 2, \text{eig} = F, \text{add} = F$) applies a mapping of the 8 environments onto a two dimensional plane based on the distances. The result is a matrix $Y = ||y_{ij}||$ where i represents the environment and j the coordinate in 2 dimensions. This matrix Y is presented in the first two columns of Table 4. The last 3 columns represent a preliminary weight w_1 indicating the importance of success in this environment and pr which is proportional to the prior probability of facing this environment, and their product which will be referred to as the weight w. The eight points are plotted in Figure 1 and circled.

On the assumption that only two tests will be permitted we select two points $\boldsymbol{x}_1 = (x_{11}, x_{12})$ and $\boldsymbol{x}_2 = (x_{21}, x_{22})$ so as to optimize a criterion. The criterion we use here to be maximized is

$$V = \min_j \left\{ \frac{1}{w_j} [I(\boldsymbol{x}_1, j) + I(\boldsymbol{x}_2, j)] \right\}$$

where $I(\boldsymbol{x}, j)$ is the information that an experiment at $\boldsymbol{x}$ contributes to the j-th environment. For this discussion let us assume that

$$I(\boldsymbol{x}, j) = \exp\left\{ -b||\boldsymbol{x} - \boldsymbol{y}_j|| \right\}$$

where $\boldsymbol{y}_j$ is the location in the two dimensional space of the point corresponding to the j-th environment. The optimizing points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and the corresponding value of V depend on the value of b. Table 5 represents the dependence on b. These points are connected in Figure 1.

ISSUES AND ALTERNATIVES

This approach is painfully lacking in adequate justification. The main reason for not dismissing it out of hand is that the underlying problem is real and demands some resolution. In this section we will review the example step by step, consider the issues raised and alternatives to the methods proposed.

The first step was the construction of a stress matrix A. Here we have ignored one of the major contributions of operational testing (OT). That consists of the illuminating surprises that accompany OT. Frankly, I don't see how to incorporate that aspect in this "model." To

construct A we have to employ enough expertise to imagine the various aspects or variables of the many environments that might impact on the quality of performance. It is necessary to quantify, in some orderly fashion, the perceived threat to satisfactory performance embodied in each of these variables. Such a quantification will almost necessarily be partly subjective and should depend at least in part on the results of developmental testing (DT). Note that in this example two of the variables were hot and cold. It might seem strange to list these as separate variables, but the stresses imposed by extremes of heat can be regarded as distinct in nature from those imposed by extremes of cold, or for that matter, of extreme shifts from cold to hot.

Implicit in the quantification of stress is that A can be used to generate a measure of distance or dissimilarity between pairs of environments. The measure of distance used here, to generate the matrix D, is naive. It might be that the expert could bypass A and go directly to D. Otherwise he might find some reasonable alternative to our definition of D. Implicitly, the definition used here weights each variable as heavily as every other and constructs a Euclidean type of distance. If some of our stress variables were highly correlated because they tend to measure the same underlying factor, our measure D could effectively give this factor more impact than other equally important factors. That phenomenon can be compensated for, if it is understood, by replacing the squared distance by some other quadratic form or by some other metric or measure altogether.

With our measure of dissimilarity, we are effectively measuring distances of points in an 18 dimensional Euclidean space. Each environment is represented by one of these points. Other measures of dissimilarity may not be able to be mapped into points in such a space. In any case, it is difficult to comprehend any analysis involving such high dimensionality. There are a number of techniques that have appeared in the statistical literature that were developed to cope with representing high dimensional phenomena in terms of a low dimensional Euclidean space. These methods go under various names and are considered to be variations of "Factor Analysis."

One of the earliest such methods is called Principle Components. This technique effectively projects a set of points in n dimensions onto the closest $k < n$ dimensional space. The meaningfulness of the result depends on the relevance of Euclidean distance in the original n-dimensional space. The classical methods of Factor Analysis involve an assumption that the data are noisy observations, the means of which are linear functions of k underlying factors. The noise on each observed data point is assumed to be independent of the noise on the others. Then there are a set of methods of "scale analysis" which tries to map a dissimilarity matrix onto a low dimensional Euclidean space so that the distances between points in the Euclidean space are close to the dissimilarities.

In practice, these methods involve getting results for low dimensions, and comparing how well they work for low values of k with the next higher value. In our example, I applied cmdscale which is a scale analysis method in Splus for $k = 2$ without taking the time to see if using $k = 3$ would be much of an improvement as measured by a "stress" criterion. Once the points are mapped into a low dimensional space, the analyst often tries to label certain directions in the k dimensional space as measuring certain underlying factors. The classical factor analysis methods come with a variety of techniques for rotating and labeling important factors. I have tended to be skeptical of these techniques, but in applied fields like psychology, the naming of these factors may be valuable and contribute to insight.

My general attitude toward all of these approaches, is that they serve a useful purpose in suggesting insights that may well be worth while pursuing systematically in other ways. These are scattergun techniques which may hit an interesting target, but are not guaranteed to work, nor to give meaningful results when results appear.

The ability to label certain directions with interpretations that make sense to the user would be of great importance for the analyst who has to communicate with a decision maker who is necessarily reluctant to decide on the basis of the output of a black box or mysterious algorithm.

Returning to our example, plotting the points on a plane (2-dimensional space) in Figure 1, we have labeled the points by the environment number and the weight. Moving diagonally from the upper left hand, one seems roughly to be moving from a temperate to an intemperate environment. Moving from left to right seems to be going from a humid to a dry environment. These characteristics are not nearly a complete description of the environments, and it is of value to keep the labels handy to remind one of the actual environment.

We would hope that the lower dimensional representation would be helpful in describing how much information an experiment in one environment gives to the user who is interested in another environment. One possibility is that an expert can be asked how much information can be obtained from an experiment in Saudi Arabia for use in a temperate urban environment and vice versa. With introspection one could conceivably construct an information matrix $I(\boldsymbol{x},\boldsymbol{y})$ representing the information from an experiment at $\boldsymbol{x}$ for use at $\boldsymbol{y}$. This matrix need not be square. We could have more or fewer values of $\boldsymbol{x}$ than of $\boldsymbol{y}$. Presumably the closer $\boldsymbol{x}$ is to $\boldsymbol{y}$ the greater the value of $I(\boldsymbol{x},\boldsymbol{y})$. Actually that need not be the case, when we consider the potential advantages of accelerated stress testing. For the time being let us defer that issue. In this example, we have assumed that $I(\boldsymbol{x},\boldsymbol{y})$ is a function of the distance from the representations of $\boldsymbol{x}$ and $\boldsymbol{y}$ in the two dimensional space on which the environments have been mapped. We have assumed that I is a decreasing function of the distance, and in particular that it can be represented by $\exp(-b||\boldsymbol{x}-\boldsymbol{y}||)$, where b is a parameter to be selected. That choice was pretty arbitrary, and it would make sense instead to ask experts their assessment of I and use that to fit some reasonable function.

The next step was to construct a criterion of what would be a good design. Here the word "design" is used to represent a choice of several experimental environments $\boldsymbol{x}$. For the sake of the example, I decided to use two experimental environments. This choice of 2 is not necessarily limited to be the same as the dimensionality of the space on which the environments were mapped. I also weighted each $\boldsymbol{x}$ equally in calculating a cumulated information by summing the informations at a given point $\boldsymbol{y}$. In practice, we may decide to spend more assets or money on one test than another. In that case we would not simply sum $I(\boldsymbol{x}_1,\boldsymbol{y})$ and $I(\boldsymbol{x}_2,\boldsymbol{y})$ where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are the two environments used for testing. We could use a weighted sum which would take into account how many assets were used as well as the environment. Thus the restrictions to treating two test locations and weighting them equally are not essential and could easily be modified. To return to the criterion, I selected that of optimizing the worst that could happen where the worst is defined as the minimum over all possible environments $\boldsymbol{y}$, of the cumulated information at $\boldsymbol{y}$ divided by the weight at $\boldsymbol{y}$. Thus a low value of information at an important environment would be much worse than a low value at the North Pole which I assumed to be of little military importance for this example.

For my example I gave positive weights only to the 8 environments used for the mapping. That need not be the case. I also considered the possibility of creating a test in any environment in the two dimensional space. That may not be feasible. We might be more limited. It may be difficult to construct an environment that would be mapped into a given point in the two dimensional space. Possibly more embarrassing, there may be too many different real environments that would correspond to a given point in the two dimensional space.

Table 5 shows the impact of changing the parameter b. When b is small, the impact of distance is slight, and it is important to make sure that the important or highly weighted points get maximum information. Then the optimal design puts the two x values at the two most highly weighted environments. When b is large, information is greatly diminished with distance from the testing point. Then it is necessary to move the x values to some compromise positions which do not downgrade too much the performance at less important places.

While this behavior may seem sensible, it depends heavily on buying into the criterion proposed. Both the exponential decline and the maximin aspects should be questioned and alternatives considered. This even ignores the possibility of replacing the information $I(\boldsymbol{x}, \boldsymbol{y})$ by a higher dimensional measure such as an information matrix. At this stage of sophistication, it seems premature to consider this latter extension.

The issue of accelerated testing has not been addressed here yet. If it were, it might be that information would not decrease as one moves $\boldsymbol{x}$ from $\boldsymbol{y}$. One possibility is to add a dimension for stress. However, the North Pole and Saudi Arabia represent high stress environments at least with respect to certain variables, and one might be able to incorporate these high stresses without going to another dimension.

HOW LARGE SHOULD THE SAMPLE BE?

We consider two examples. the first describes some of the consequences of selecting a given sample size in establishing a rather modest 80% confidence statement on the unknown reliability of a system. The second deals with finding the appropriate sample size for testing whether the mean of a normal distribution exceeds a desired threshold value. A special characteristic of the latter problem is that the sample size must be selected in advance of experimentation. A subsequent criticism suggests the potential benefit of a multiple or sequential sampling plan.

CONFIDENCE BOUNDS ON RELIABILITY FOR A GIVEN SAMPLE SIZE

Some of the underlying problems with deciding sample size to determine reliability can be understood in terms of the following table dedicated to providing 80% confidence that a given reliability is at least 80%. Suppose that n items are tested in a success failure mode to determine whether there is 80% confidence that it has reliability at least 80%. Let s be the number of successes required to *pass the test.*

First we note that to pass the test, n must be at least 8. Let r be the actual reliability required so that we will be 80% sure of passing this test. Also let q be the reliability below which the probability of passing the test is less than 0.10. Finally let t be the probability of passing the test when the reliability is exactly 80%.

The irregularities in the trends in Table 6 derive from the discrete nature of the binomial distribution.

In any case a very high reliability r is required to be moderately sure of passing the test for $n \leq 50$. Also the probability of passing the test decreases rapidly as the reliability decreases.

A TESTING PROBLEM

While operational testing is not often presented as a pass fail test of a system, the problem we pose here is exactly that. Although it is phrased in terms of normally distributed observations, the illustrative example will involve a binomial problem. The same normal problem has been solved by Grundy et al.[1] in a slightly different context. It was also discussed by Raiffa and Schlaifer[2]. The presentation of the results here differ from those in the other publications, and fits in better with a sequential version of this problem, a major topic of Chernoff[3].

Let $X_1, X_2, \ldots, X_n$ be independent identically distributed normal random variables with unknown mean μ and known variance σ^2. The unknown parameter μ has the prior normal distribution $N(\mu_0, \sigma_0^2)$ with known mean μ_0 and known variance σ_0^2. It is desired to decide whether $\mu > 0$ or $\mu < 0$, and the cost of an incorrect decision is $k|\mu|$ where k is known. The cost of n observations is cn where c is known. How large should n be?

POSTERIOR DISTRIBUTION AND RISK

The posterior distribution of μ given the data $X_1, X_2, \ldots, X_n$ is $N(Y, s)$ where

$$Y = (\mu_0\sigma_0^{-2} + n\sigma^{-2}\overline{X})/(\sigma_0^{-2} + n\sigma^{-2}) \tag{1}$$

is the *Bayes posterior estimate* of the unknown mean $\mu, \overline{X}$ is the average of the n observations, and

$$s^{-1} = (\sigma_0^{-2} + n\sigma^{-2}) \tag{2}$$

is the *precision* of the posterior estimate Y of μ. Note that Y is a weighted average of the prior mean μ_0 and the average $\overline{X}$ weighted by their precisions σ_0^{-2} and $n\sigma^{-2}$. In a sense the prior distribution corresponds to information gathered from

$$n_0 = \sigma^2/\sigma_0^2$$

observations averaging μ_0.

The appropriate decision rule for this symmetric problem is to decide $\mu > 0$ if and only if $Y > 0$. The *posterior risk* associated with this procedure is

$$r_n(Y) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi s}} e^{-\frac{1}{2s}(\mu - Y)^2} k|\mu| d\mu + cn$$

if $Y > 0$. In general, $r_n(Y)$ is an even function, and

$$r_n(Y) = ks^{1/2}\rho(\alpha) + cn \tag{3}$$

where

$$\alpha = Y/\sqrt{s}$$

is the number of standard deviations of the posterior estimate Y from 0,

$$\rho(\alpha) = \phi(\alpha) - |\alpha|\{1 - \Phi(|\alpha|)\}$$

and ϕ and Φ are the standard normal density and cumulative distribution functions, i.e.,

$$\phi(\alpha) = (2\pi)^{-1/2} \exp(-\alpha^2/2)$$

and

$$\Phi(\alpha) = \int_{-\infty}^{\alpha} \phi(v) dv.$$

Before the data are observed to yield a posterior risk, the *Bayes risk* is the expectation

$$R_n = E\{r_n(Y)\} = \int_{-\infty}^{\infty} r_n(\mu_0 + \epsilon\sqrt{s_0 - s})\phi(\epsilon) d\epsilon \tag{4}$$

where $s_0 = \sigma_0^2$. Then, it can be shown that

$$R_n = k\{s_0^{1/2}\rho(\alpha_0) - s_1^{1/2}\rho(\alpha_1)\} + c\sigma^2 s^{-1} - c\sigma^2/\sigma_0^2 \tag{5}$$

where $\alpha_0 = \mu_0/\sqrt{s_0}$, $\alpha_1 = \mu_0/\sqrt{s_1}$ and $s_1 = s_0 - s$.

NORMALIZATION

Here R_n depends on the known constants $\mu_0, \sigma_0^2, \sigma^2, k$ and c. It is obvious that the number of effective parameters for describing the optimal choice of n can be reduced. For example, c/k is, except for a trivial normalization, more relevant than c and k separately.

A valuable normalization reducing the number of effective parameters is obtained as follows. Let $\tilde{X}_i = aX_i$. Then $\tilde{\mu} = a\mu, \tilde{\sigma}^2 = a^2\sigma^2, \tilde{\mu}_0 = a\mu_0, \tilde{\sigma}_0^2 = a^2\sigma_0^2, \tilde{Y} = aY, \tilde{s} = a^2 s, \tilde{s}_0 = a^2 s_0, \tilde{s}_1 = a^2 s_1, \tilde{\alpha} = \alpha, \tilde{\alpha}_0 = \alpha_0$ and $\tilde{\alpha}_1 = \alpha_1$. Next

$$R_n = ka^{-1}\{\tilde{s}_0^{1/2}\rho(\tilde{\alpha}_0) - \tilde{s}_1^{1/2}\rho(\tilde{\alpha}_1)\} + c\sigma^2 a^2 \tilde{s}^{-1} - c\sigma^2/\sigma_0^2,$$

and selecting a to make $ka^{-1} = c\sigma^2 a^2$, i.e.,

$$a = k^{1/3} c^{-1/3} \sigma^{-2/3} ,$$

we have

$$k^{-2/3} c^{-1/3} \sigma^{-2/3} [R_n + c\sigma^2/\sigma_0^2] = \tilde{s}_0^{1/2}\rho(\tilde{\alpha}_0) - \tilde{s}_1^{1/2}\rho(\tilde{\alpha}_1) + \tilde{s}^{-1}.$$

Thus the choice of the optimal sample size is essentially that of minimizing

$$\tilde{R} = \tilde{s}_0^{1/2}\rho(\tilde{\alpha}_0) - \tilde{s}_1^{1/2}\rho(\tilde{\alpha}_1) + \tilde{s}^{-1} \tag{6}$$

with respect to $\tilde{s}$. We recall that $\tilde{s}_0$ and $\tilde{\alpha}_0 = \tilde{\mu}_0\tilde{s}_0^{-1/2}$ are fixed, and $\tilde{s}_1 = \tilde{s}_0 - \tilde{s}$ and $\tilde{\alpha}_1 = \tilde{\mu}_0\tilde{s}_1^{-1/2}$ depend on $\tilde{s}$. Setting $d\tilde{R}/d\tilde{s}$ equal to 0 yields

$$\frac{1}{2\tilde{s}_1^{1/2}}\phi(\tilde{\alpha}_1) = \tilde{s}^{-2} \tag{7}$$

from which we can derive level lines for the optimal

$$\tilde{n} = \tilde{s}^{-1} - \tilde{s}_0^{-1}. \tag{8}$$

This result then easily gives the optimal

$$n = \sigma^2(s^{-1} - \sigma_0^{-2}) = (\sigma k/c)^{2/3}\tilde{n} \tag{9}$$

as a multiple of $\tilde{n}$ for each value of $(\tilde{\mu}_0, \tilde{\sigma}_0)$ or, equivalently, of $(\tilde{t}_0, \tilde{\alpha}_0)$ where

$$\tilde{t}_0 = \tilde{s}_0^{-1} = \tilde{\sigma}_0^{-2} = (\sigma k/c)^{-2/3}(\sigma^2/\sigma_0^2) \tag{10}$$

and

$$\tilde{\alpha}_0 = \alpha_0 = \mu_0/\sigma_0$$

Because of the normalization, $\tilde{n}$ assumes fractional values. The discrete nature of our original problem implies that the appropriate value of n is typically a nearby integer value of $n > 0$, unless $\tilde{n} = 0$ in which case $n = 0$.

RESULTS

The results of the calculation of the optimal choice of $\tilde{n}$ are presented graphically in Figure 2 which shows the curves in the $(\tilde{t}_0, \alpha_0)$ plane for which $\tilde{n}$ takes on the values .02, .04, .06, .08, 1.0, 1.25, 1.50 and 1.75. Because of symmetry, Figure 2 is given only for $\alpha_0 \geq 0$. Some consequences of these results are worth emphasizing. It does not pay to sample if $\tilde{t}_0$ is too large. If $\tilde{t}_0$ is too small then a minimal amount of sampling, i.e., $n = 1$ is required. Given that $t_0 = \sigma_0^{-2}$ is a measure of the prior precision or information, it is not surprising that large values of t_0 should discourage sampling, especially if μ_0 is large. However, it may seem paradoxical that we should not wish to sample much when we are almost completely ignorant, *a priori*, about μ.

The explanation, from a Bayesian perspective is that when t_0 is small, μ is very unlikely to be moderate in size. Thus one observation will be enough to determine whether μ is highly positive or highly negative. Even if $\mu_0 = \alpha_0 = 0$, a single observation should suffice as long as we have prior reason to believe that $|\mu|$ is large compared to σ.

The value of $\tilde{n}$ can never exceed 1.8064 which it attains when $\alpha_0 = 0$ and $\tilde{t}_0 = 0.0904$.

There is a boundary of values $(\tilde{t}_0, \alpha_0)$ for which $\tilde{n} = 0$. On this boundary, there is a sample size $\tilde{n} > 0$ which gives the same Bayes risk as $\tilde{n} = 0$. The underlying reason for this *bifurcation* effect is that the risk, as a function of increasing $\tilde{n}$, as $\tilde{t}_0$ and α_0 are kept fixed, first increases and then decreases to a minimum value before going to ∞. When the minimum value of the risk is below that for $\tilde{n} = 0$, it pays to sample. As we fix α_0 and change $\tilde{t}_0$, the location of the minimizing value of $\tilde{n}$ and $\tilde{R}$ change gradually and the minimizing value of $\tilde{n}$ approaches a positive limit as the minimum value $\tilde{R}$ approaches the risk corresponding to $\tilde{n} = 0$.

BINOMIAL ILLUSTRATION

I propose to illustrate the solution with an artificial binomial example involving a missile which either succeeds or fails. In principle, a binomial problem can be solved directly and our normal approximation to that problem is unnecessary. However it serves a useful illustrative purpose.

Even in a relatively simple binomial problem where the prior distribution of the unknown probability of success, p, has a beta distribution, it isn't possible to find a normalization that

reduces the problem to a graph comparable to Figure 2. Figure 2 is useful for quick and dirty answers, and overall insight. Of course, precise results for specific problems deserve more detailed analysis, especially when large amounts of money are involved. If small sample sizes are called for, then the normal approximation may be unreliable.

Example. It is proposed to design a new missile to upgrade the reliability from the current value of 70% to 85% at a cost of 10 billion dollars to produce 1000 missiles. If the new missile achieves a reliability p of only 80%, the effort will have seemed barely worthwhile. Thus we estimate $k = 1$ representing the cost in billions of dollars per percentage deviation from 80. The cost per missile tested is 10 million dollars or $c = .01$ in billions of dollars. Assuming that $100p = \mu + 80$, the observations $X = 100$ on success and 0 on failure have standard deviation $\sigma = 100\sqrt{.2 \times .8} = 40$. The engineers feel that they will reach 85% reliability and the prior on μ has mean $\mu_0 = 5$ and standard deviation $\sigma_0 = 8$ (equivalent to 25 observations).

Then $a = (k/c\sigma^2)^{1/3} = 0.39685$, $\tilde{t}_0 = (\sigma k/c)^{-2/3}(\sigma^2/\sigma_0^2) = 0.09921$ and $\alpha_0 = \mu_0/\sigma_0 = 0.625$.

The corresponding value of $\tilde{n}$ is 0.1431 and the optimal $n = \sigma^2 a^2 \tilde{n} = 36.06$ which can be rounded off to 36 costing 360 million dollars for testing missiles, not counting the set up cost.

RATIONALE

Two aspects of the problem require some rationalization. These are the $k|\mu|$ cost for wrong decision and the use of the normal distributions. Incidentally, the linear cost of sampling makes sense even if there is a set up cost, providing the cost of additional observations are approximately fixed per observation.

Suppose that the appropriate decision depends on the size of some unknown parameter μ. We wish to make one decision if the parameter is large and an alternative decision if it is small. Generally that means that the loss (or payoff) for each decision is some function of the parameter. These two functions intersect at some break even point μ^* of μ. If these two functions are differentiable at μ^*, the difference between the two functions is approximately $k(\mu - \mu^*)$ for μ near μ^* where k is the derivative at μ^* of the difference between these two functions. The loss for taking the wrong action is then approximately $|k(\mu - \mu^*)|$. By translating the parameter from μ to $\mu - \mu^*$, we have a break even point at 0 and a loss of $k|\mu|$ where k is positive.

The illustrative example involving the binomial shows how we can approximate other problems by the normal problem when we can rely on the central limit theorem. If exact results are called for, then this normal problem should be regarded as a convenient means of obtaining a rough approximation.

SUMMARY

The nature of the solution which calls for minimal sampling when little is known about μ was partly explained in the section on results. If indeed, we are almost certain $|\mu|$ is large, then that solution is appropriate. That resolution is unsatisfying to a frequentist who would like a more robust solution which would be likely to lead to a low expected loss no matter what the value of μ is. The frequentist might prefer a minimax procedure which would call for $n = 0.1933(\sigma k/c)^{2/3}$ observations. In that case the maximal risk is attained at $\mu = 0.7518\sigma/\sqrt{n}$. In the binomial illustration, this approximation would have $n = 48.7$.

Both the Bayesian and the frequentist are inclined to dislike the severe restraint that the sample size must be determined in advance before any testing is carried out. The potential advantages of sequential sampling, or at least double or triple sampling, should not be neglected, especially in those cases where the prior distribution is very vague. The relaxation of the above restraint will help considerably to ameliorate the difficulty.

In case we wish to consider further sampling, after n observations have been made. Figure 2 is still of value. The data have served to convert our prior α_0 to a posterior $\alpha = Y/\sqrt{s}$ and $\tilde{t}_0$ to $\tilde{t} = \tilde{s}^{-1} = \tilde{t}_0 + \tilde{n}$. Replacing $(\tilde{t}_0, \alpha_0)$ by $(\tilde{t}, \alpha)$ we may use the figure to decide how large an additional sample size should be, if only one more such choice is allowed. While Figure 2 is useful in telling us how large the second sample should be in a two sample case it does not help to tell us how large the first sample should be in a two sample study.

In the two sample case we should expect the first sample size to be relatively small compared to n when $n > 0$. However the option of taking a small first sample should extend the $(\tilde{t}_0, \alpha_0)$ range over which we should do some sampling. If we should decide to proceed in a fully sequential mode, deciding after each observation whether or not to continue sampling, then the appropriate sequential stopping boundary is given by the dashed curve in Figure 2 to the right of the curves describing $\tilde{n}$. In the sequential case the proper labeling of the axes would be $\tilde{t}$ and α. Sampling should continue as long as $(\tilde{t}, \alpha)$ is to the left of the dashed curve.

In summary, the normal problem has the advantage of the normalization that makes the two dimensional Figure 2 useful. The representation in terms of $(\tilde{t}, \alpha)$ is useful in two ways. First $\tilde{t}$ measures the cumulated precision or information. Second, since α measures the number of standard deviations from 0, it provides a nominal significance level. For example $\alpha = 1.5$ corresponds to the nominal significance of $\Phi(-1.5) = 0.067$.

REFERENCES

1. Grundy, P.M., Healy, M.J. R. and Rees, D.H. Economic Choice of the Amount of Experimentation, J. Roy. Statist. Soc. Ser. B, Vol. 18, pp. 22-49, 1956.

2. Raiffa, H. and Schlaifer, R. Applied Statistical Decision Theory, Division of Research, Harvard Business School, Boston, MA, 1961.

3. Chernoff, H. Sequential Analysis and Optimal Design, SIAM, Philadelphia, PA., 1972.

Table 1 Stresses in Various Environments, $A = ||a_{ij}||$

i/j	1	2	3	4	5	6	7	8
1	9	8	1	1	5	6	4	6
2	2	1	9	8	5	6	7	5
3	4	2	2	3	5	7	8	5
4	9	1	8	8	5	6	7	5
5	1	9	2	2	5	4	3	6
6	3	3	3	3	5	7	7	6
7	7	3	1	1	5	6	4	6
8	3	5	1	1	5	7	5	6
9	7	4	7	9	5	7	8	4
10	7	5	8	10	5	7	8	5
11	2	2	2	10	2	6	8	3
12	8	5	5	5	5	5	4	3
13	2	5	2	2	5	6	7	7
14	3	4	3	3	5	6	6	8
15	8	3	3	7	8	6	3	9
16	8	2	5	8	8	7	7	9
17	7	5	4	4	5	5	3	8
18	7	4	7	7	5	4	6	4

Table 2 Stress Variables (Rows) and Environments (Columns) of Table 1

Rows

temperature
1. hot
2. cold
3. variability

humidity
4. dry
5. humid
6. variability

dust
7. particle size
8. standard dev. of part. size
9. windiness
10. peaks of windiness

altitude
11. altitude

demand
12. heavy
13. irregular
14. peaks

fuel
15. available
16. quality

service
17. parts available
18. quality of personnel

Columns

1. Saudi Arabia
2. jungle
3. North Pole
4. Himalaya peak
5. temperate rural ground level
6. temperate rural hilly
7. temperate rural mountain
8. temperate urban

Table 3 Distance Matrix $D = ||d_{jj'}||$

i/j	1	2	3	4	5	6	7	8
1	0.00	16.12	14.56	15.46	10.39	12.37	15.30	13.42
2	16.12	0.00	16.67	20.07	11.75	14.18	16.73	13.86
3	14.56	16.67	0.00	9.95	12.65	14.46	13.04	16.85
4	15.46	20.07	9.95	0.00	14.39	14.00	11.87	17.46
5	10.39	11.75	12.65	14.39	0.00	7.00	10.86	6.00
6	12.37	14.18	14.46	14.00	7.00	0.00	6.40	8.06
7	15.30	16.73	13.04	11.87	10.86	6.40	0.00	12.65
8	13.42	13.86	16.85	17.46	6.00	8.06	12.65	0.00

Table 4 Result of cmdscale $(D, k = 2, \text{eig} = F, \text{add} = F)$ and weights

$$Y = ||y_{ij}||$$

i/j	1	2	w_1	pr	w
1	-0.89	-2.47	3	5	15
2	-7.88	-6.95	2	3	6
3	7.03	-5.25	1	1	1
4	10.06	-0.08	1	1	1
5	-3.22	0.52	5	5	25
6	-1.81	4.90	3	5	15
7	2.80	5.40	1	2	2
8	-6.08	3.93	5	5	25

The rows represent the environments.
The first two columns are the coordinates in the two-dimensional space.
The next three columns are the weights representing importance, the prior probabilities that these environments will show up, and the product of those two. Note that the prior probabilities are not normalized to add to one. Also the weight is labeled w_1 while the product is labeled w because it will hereafter be referred to as the weight.

Table 5 Optimizing Points and Values for Different Values of b

b	x_1		x_2		v
	x_{11}	x_{12}	x_{21}	x_{22}	
.00	-3.22	0.52	-6.08	3.93	8.0(-2)
.20	-2.88	-0.33	-5.83	3.65	5.1(-2)
.30	-2.74	-2.80	-4.83	3.17	3.0(-2)
.40	-0.87	-3.85	-4.58	1.24	1.3(-2)
.50	-0.17	-0.12	-6.74	-0.13	6.2(-3)
.60	0.65	1.45	-6.78	-0.45	3.3(-3)
.80	1.59	1.42	-6.84	-0.67	1.0(-3)
1.00	2.12	1.11	-6.42	-0.88	3.3(-4)
1.50	2.53	0.86	-4.90	-1.29	1.1(-5)
3.00	2.92	0.63	-4.79	-1.14	+

Table 6

n	s	r	q	t
8	8	0.973	0.750	0.168
10	10	0.978	0.795	0.107
25	22	0.907	0.752	0.234
50	42	0.869	0.754	0.307
100	83	0.854	0.772	0.271
400	327	0.832	0.790	0.209

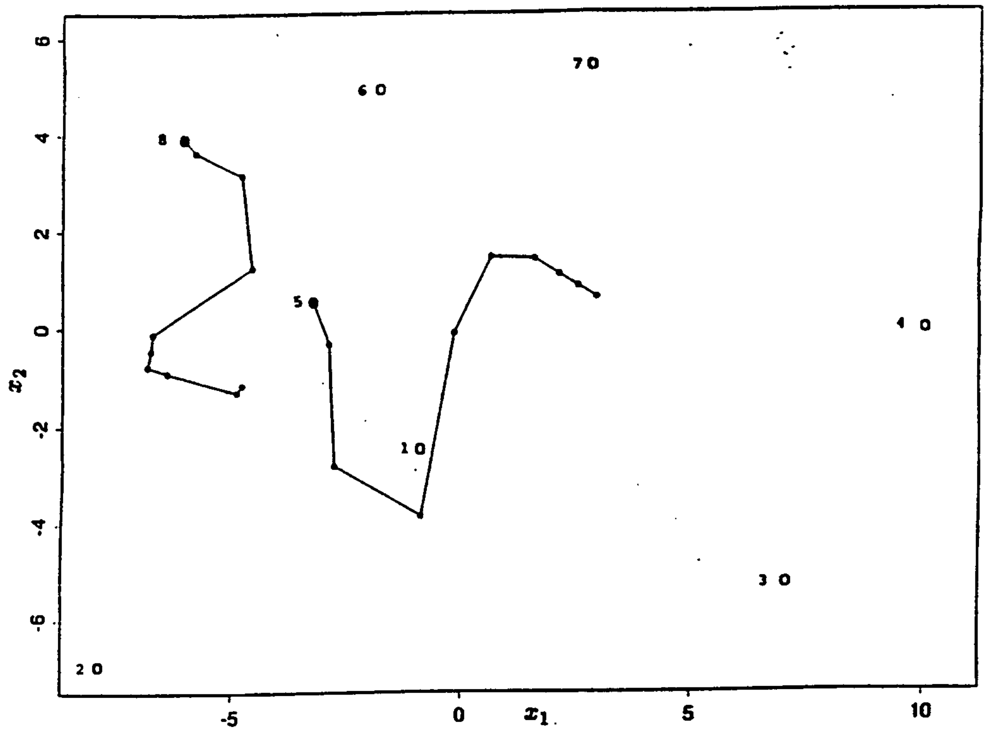


Figure 1: Location of Environments and Optimal Design Points

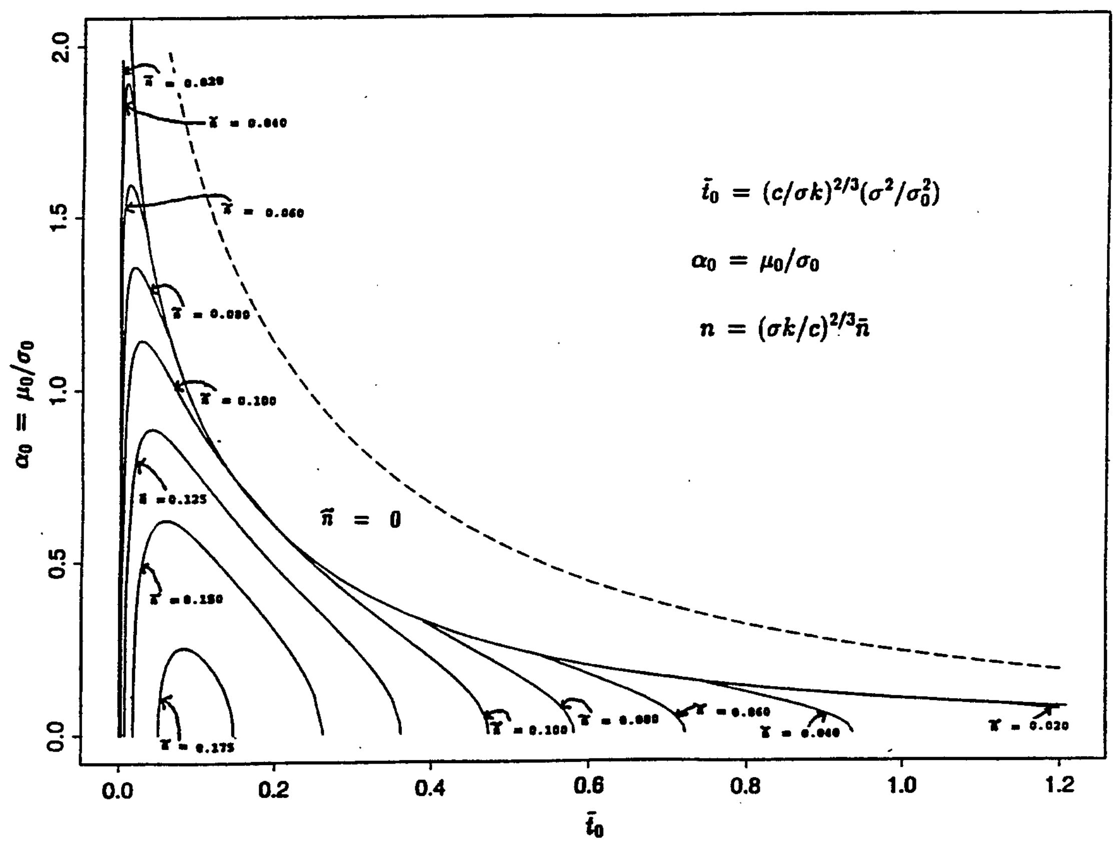


Figure 2: Level Lines of Optimal Sample Size, Normalized

INTENTIONALLY LEFT BLANK.

OVERVIEW OF EXPERIMENTATION AT FORT HUNTER LIGGETT

(Introduction to Special Session on *"Forty Years of Experimentation at Fort Hunter Liggett"*)

Carl T. Russell
Chief Scientist, TEXCOM Experimentation Center
Fort Hunter Liggett, California 93928-8000

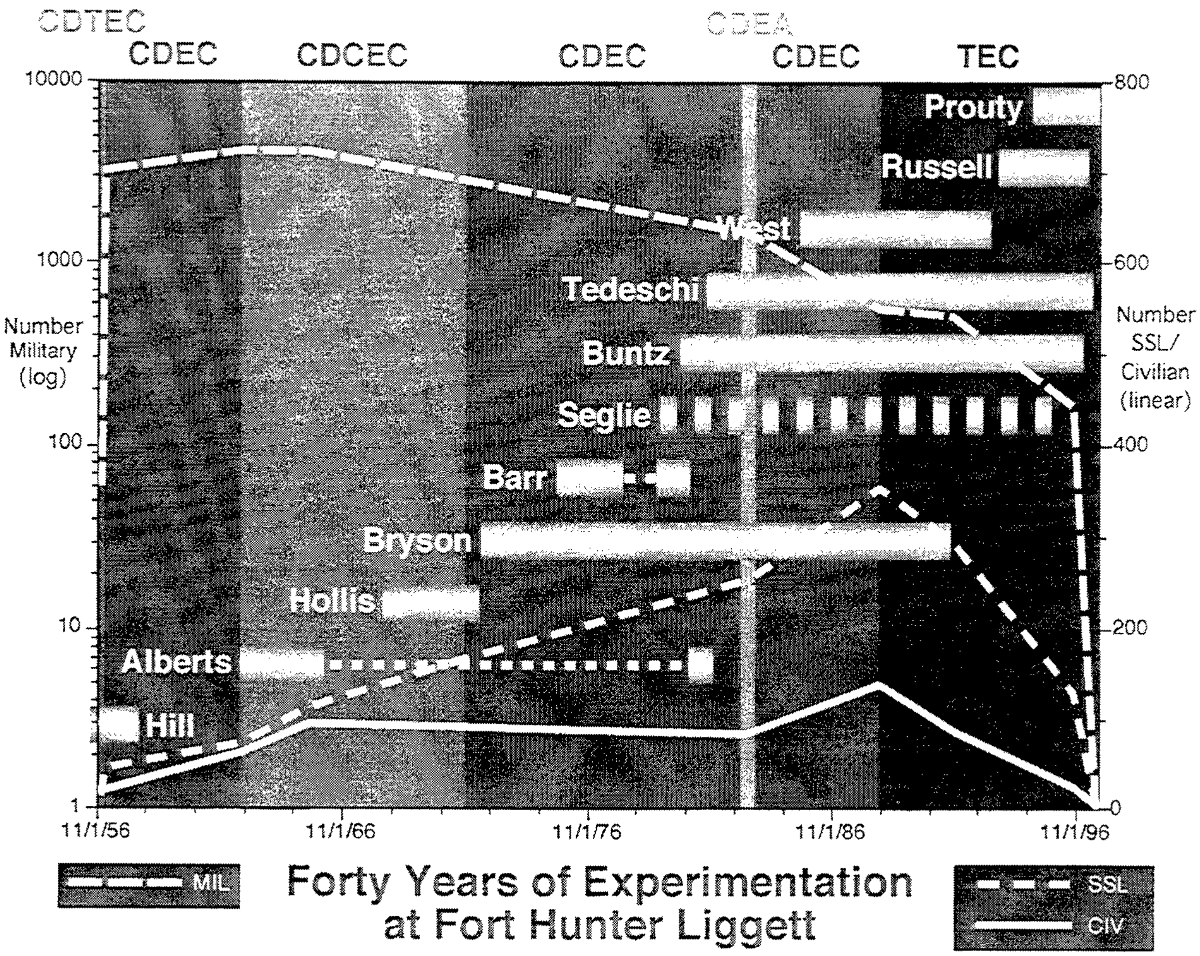
TEXCOM Experimentation Center began its life as the Combat Developments Test and Experimentation Center on 1 November 1956 and first became Combat Development Experimentation Center (CDEC) on 1 January 1957. Subsequent history is dominated by the name "CDEC" as is clear from the following table.

CDTEC	Combat Developments Test and Experimentation Center	11/01/56-12/31/56
CDEC	Combat Development Experimentation Center	01/01/57-06/30/62
CDEC	Combat Developments Experimentation Center	07/01/62-04/30/63
CDCEC	Combat Developments Command Experimentation Center	05/01/63-03/22/65
CDCEC	Combat Developments Command Experimentation Command	03/23/65-08/31/71
CDEC	Combat Developments Experimentation Command	09/01/71-03/22/83
CDEA	Combat Developments Experimentation Activity	03/23/83-07/01/83
CDEC	Combat Developments Experimentation Center	07/02/83-11/02/88
TEC	TRADOC Test and Experimentation Command, Experimentation Center	11/03/88-11/14/90
TEC	Test and Experimentation Command, Experimentation Center	11/15/90-09/30/97

The name has always contained "Experimentation," and until 1988 it always started with "Combat Developments." This is important because CDEC was established expressly for experimenting with organizational concepts as well as doctrinal and materiel concepts. As such it had no predecessor and no existent body of experimental method—that had to be learned and developed from scratch during the early days. From the beginning, experimentation at Fort Hunter Liggett has concentrated on performing Real-Time Casualty Assessment (RTCA) in force-on-force trials to make those trials replicate combat as closely as possible and on measuring the results of those trials as accurately as possible—and instrumentation has always played a prominent part. From the late 1970's onward, CDEC's workload became more-and-more oriented towards operational testing, partly accounting for the name change in 1988. With increased spending constraints in DoD, the Army has determined that maintaining an experimentation facility at Fort Hunter is no longer affordable, so the Command will inactivate effective 30 September 1997.

Although they have historical content, the talks in this Special Session are primarily about harvesting and interpreting data for making important decisions. The talks by no means fully document the technical history of CDEC, but as a varied series of vignettes they sketch CDEC's role in Army experimentation over the past forty years.

Most technical support for CDEC has always been in the hands of a government contractor. In the earliest years, this contractor worked directly for the Commanding General. Until 1966 when the contractor became known as the Scientific Support Laboratory (SSL), the contractor was know as the "Research Office." **Floyd Hill** was hired in the summer of 1956 to start staff recruitment and begin program planning for CDEC. By November 1st he had office space in Monterey and a staf of ten professionals to continue planning and methodology development for the first experiment at CDEC in March 1957. When CDEC was made permanent in 1958, Stanford Research Institute won the competition. Increasing requirements for new instrumentation to meet CDEC data needs continued, and from 1962 to 1966 much instrumentation was developed and fielded under the direction of **Henry Alberts** who headed SSL instrumentation again in 1980-81.



Forty Years of Experimentation at Fort Hunter Liggett

In July, 1968, **Walter Hollis** was appointed to the newly-created position of "Scientific Advisor," at CDEC, supplementing technical advice to the Commanding General from the contractor side with advice from the government side. During his tenure at CDEC, experimentation transitioned from stopwatch to computer as instrumentation and automation capabilities advanced. **Marion Bryson** replaced Mr. Hollis as Scientific Advisor in 1973, and he became the Director of CDEC in 1983, remaining until he left to become TEXCOM Technical Director in 1991. As you can imagine, much changed during his tenure at CDEC. **Bill West** came to CDEC in 1985 as Chief Scientist under Dr. Bryson and remained after Dr. Bryson left as Chief Scientist and Deputy Director. **Carl Russell** replaced Bill West as Chief Scientist in 1993. **James Prouty** has been the TEC Commander since August, 1995, but COL Prouty spent substantial time at CDEC earlier in connection with TASVAL and Apache Hellfire testing.

The current TEXCOM Technical Director, **Brian Barr**, was assigned to CDEC as a Captain in 1975-78, and he returned as a Major in 1979-80. His paper will address some classic data from the early 1970's. In 1979 as a new IDA employee fresh from teaching physics at Yale, **Ernest Seglie** cut his DoD teeth on TASVAL, and he has returned to Fort Hunter Liggett many times since, mostly in his oversight role as Scientific Advisor to the Director of Operational Test and Evaluation. He is the only person shown on this slide who was never assigned to CDEC. Dr. Seglie initiated the National Research Council project which **Herman Chernoff** discussed this morning, and his paper this afternoon will assess the importance of high-resolution RTCA in operational testing. **Ed Buntz**, the current instrumentation chief, came to CDEC as a Captain in 1980, was promoted to Major here, and like many others who ended their military career at CDEC, he never left. **Mike Tedeschi**, chief of methodology under Mr. Buntz, joined CDEC in 1981. Together, they have led an effort which not only made RTCA instrumentation mobile but also produced what is arguably the Army's best After Action Review capability.

At the final session of the day, Mr. Barr will moderate a panel in which Mr. Hollis, Dr. Seglie and Dr. Bryson discuss the future of field experimentation.

RECOLLECTIONS OF FIRST YEARS OF CDEC - SEPTEMBER 1956 TO JUNE 1958

Floyd I. Hill
Associate Director, Research Office Experimentation Center

ABSTRACT

Because the documentation for these years was largely destroyed, the paper is entitled "Recollections." Mr. Hill was hired because of his experience in designing and directing the first Operational Test, Project STALK, conducted at Fort Irwin August–December 1953. This test is briefly outlined, and its influence on the first CDEC experiments frequently noted. The first major supported company-sized organizational experiment using the two-sided operational game as a model is described in terms of instrumentation and umpire procedures. Some findings are given. The reasons underlying the breakdown of the contractual relationship between General Fred Gibb, Commanding General of CDEC and the Research Office are given. Some of the subsequent tests, including the Scouting Experiment and the Helicopter Experiment, are discussed. At no place did the results of these operational tests (including STALK) conform to existing policy. They were, therefore, rejected.

OPENING

Thank you for coming. My special thanks are given to Carl Russell, who, in his invitation call, speculated from his conversations with me over the years that Project STALK was the first CDEC experiment. Note that, due to document destruction, the available record of those times is at best fragmentary and often sometimes scrambled. These are my recollections, which are, no doubt, selective in nature.

COMMAND IMPLEMENTATION

The CG of TRADOC (which included the Combat Development Command [CDC] at that time), General Willard Wyman had assigned 34 senior staff offices to judge Reorganized Current Infantry Division (ROCID) field exercises in the Spring of 1956. The disturbing result was that 17 of the 34, judged ROCID to be very good and 17 judged it to be very bad. General Wyman decided that he needed objective information. So, with no appropriated R&D funds, he directed that funding be from O&M funds; troop support and CDEC staff housing be provided by the 7th Infantry Division at Fort Ord; training and testing areas be at Camp Roberts and Hunter-Liggett Military Reservation (HLMR); and the TRADOC resident scientific support contractor, Technical Operations Inc. (TOI) staff and house a 20-man scientific support group to be "Headed by a Ph.D." In September 1956 I was hired by TOI's President, Dr. Fred Henriques, to be Associate Director of the planned Research Office of the Experimentation Center (ROEC), and I was given the task of building the staff. I was the only person on the ROEC staff.

WHY I WAS HIRED (PROJECT STALK)

Dr. Henriques hired me because I had been Technical Director of Project STALK, a joint Ballistic Research Laboratories/CDC comparative test run at Camp Irwin, California, of 11 different tank/fire control systems supplied on 5 tanks, the Baseline M4A3E8, the (Light) T41, and the (Medium) M47, M47E1 (Stabilized) and T48. The effectiveness measure of the test was: time from target acquisition to a hit on a suddenly appearing target to the tank/fire control candidate on 5 (6 feet by 6 feet) stationary targets placed within 90° to either side of the tank axis of travel along a trail about 2,500 yards long. The five targets were distribution in 4 500-yard range brackets between 250 and 2,250 yards. The experimental design was an 11 × 11 Graeco Latin Square-treating the 4 main effects of: Tank/fire control system combination, Tank Crew, Test Course, and Order of Crew Testing. It was the first Army Operational Test and the only one (that I know of) that measured the effect of Player Uncertainty of where and when the target would appear by comparing the hitting performance of each tank/fire control system on a single Training Test Course to that on the 11 Record Courses, which were traversed only once by each tank crew. 25 tank crews were trained in 5 platoons for the first phase of testing and rotated, trained on, and fired a different tank in the next phase. There were 5 phases in all. 13,000 main gun rounds were fired in all. Firing at each target continued until a hit was obtained. While no problem on the Training Test Course, fired 11 times by each crew, detection of the Day-Glo paper marked targets on the Record Courses was a problem. Over one-half of the targets had to be pointed out to the Tank Commander by the Tank Controller (after the target had already been in view for 200 yards of tank travel). The joint sponsor of Project STALK (CDC) and nearly all the R&D community rejected the results of the tests, despite their extensive coverage by numerous observers, who found little to fault. The results were deemed as "Too Controversial." The most "controversial" results were:

ONE: The Baseline M4A3E8 achieved the fastest time from target acquisition to hit on the Record Courses irrespective of range or other main effects.

TWO: The T48 with the range finder/ballistic computer fire control combination was the slowest of all the tank/fire control systems on the 11 Record Courses. The newest and "best" was worst.

THREE: The foregoing results were nearly reversed on the Training Test Course which was fired 11 times by each tank crew.

FOUR: First round hitting probability (the R&D community's conventional measure) and the time to hit were simply not correlated.

Despite the Project STALK results, I was hired by Dr. Henriques based on pressure from some source. I have no idea of the source. Almost certainly the pressure to hire passed through General Wyman. I quit my job at Operations Research Inc. and commuted from Washington to TRADOC at Fort Monroe, VA.

DIRECTION AND STAFFING

When CDEC opened 1 November 1956 at Fort Ord, Brigadier General Fred Gibb and I were faced with General Wyman's command directive to complete the first step of ROCID testing by 1 July 1957, when California reserve troop training would begin at HLMR. General Gibb had a staff of 36 senior combat experienced officers. None had any test experience. Before Christmas, the ROEC staff was assembled. It consisted of new hires by me except for Dr. Frank Brooks, the Director, supplied from the TOI TRADOC staff. In all, there were 7 Ph.D.'s, and 9 MS+ (including me). Only 4 were BS-level; including two high speed computer specialists and one person with Project STALK experience, George Scott, whom I had first hired at BRL. There was one post-WWII retired Army Armored Officer, Colonel Wesley W. Yale, whom I had interviewed at the strong urging of General Wyman. Colonel Yale was one of the smartest men I have ever known, and he had a superb knowledge of strategy, tactics, terrain, and the Pre-WWII Army organizational and planning exercises.

PLANNING THE FIRST FIELD EXPERIMENT

Not surprisingly, nearly all of the elements of the January 1957 Outline Test Plan were prepared in the ROEC offices with a strong CDEC staff representation. Fort Ord was still modifying barracks to accommodate CDEC Headquarters. The resulting Plan included a strong dose of Project STALK and Colonel Yale expertise. It included:

ONE: The proposal to test 4 alternative ROCID company-level candidate organizations. These were called Integrated Combat Group (ICG) Company Organizations. Each ICG candidate organization would be tested by all of the four "friendly" companies assigned. One "aggressor" motorized Company augmented by tanks, antitank weapons, machine guns, and mortars would remain the same throughout the trials. Each ICG Company Group was a ROCID company alternative augmented by tanks, antitank weapons, and mortars. Both the "aggressor" and the "friendly" companies would be supported by artillery. Both forces would be given a mission assignment and tactical boundaries with relatively free-play in the course of 5 interconnected trials over terrain not previously operated on by the "friendly" (ROCID candidate) company personnel.

TWO: The test would use a 4 × 4 Graeco-Latin Square experimental design to treat the main effects of ROCID company organization; "friendly" company personnel; combat terrain and situation; and order of testing. There would be 4 Record Courses and a separate Training Test Course at HLMR. Training and retraining of the "friendly" companies would be at Camp Roberts.

THREE: The data record would include the: space-time, response-time, and target characteristics of the opposing forces, as well as casualty assessment and deletion from play. 2 umpire companies were requested from Fort Ord to be trained to report this information to a Master Control Station and to follow its casualty ascriptions in designating specific casualties. These umpire/controllers were to be assigned to each squad, tank, and antitank weapon, as well

as platoon and company leadership of friendly and hostile forces. They would be trained on the Record Courses and the Training Test Course.

FOUR: The test scenarios (4 for the Record Courses and 1 for the Training Test Course) were arranged so that: tactical moves would be mainly along the wooded ridges at HLMR; mission objectives were primarily on high ground; and the open valleys would be crossed, but they would be used as advance or withdrawal routes as little as possible. All 5 mission objectives for each trial were the attack or defense of a specified piece of terrain. Colonel Wesley Yale was a dominant force in scenario design. It also must not be forgotten that the 2nd Division's experience in the withdrawal from the Yalu River was still fresh in their minds. The new Army slogan—"We will not fight for real estate"—meant little operationally to these combat experienced officers.

FIVE: The measure of the relative effectiveness of the ROCID candidates would be some combination of: Enemy casualties, Friendly casualties, and Time of Mission Accomplishment. Many of the ROEC staff wrestled with a single combination and expression, but came to the conclusion that it might be used as a three-dimensional vector. More about this later.

INSTRUMENTATION, COMMUNICATIONS, AND CONTROL

Until nearly the end of the rainy season, CDEC and ROEC wrestled with instrumentation, communications, and control. In addition CDEC was heavily involved in administration of training of the 7th Infantry Division units that were all drawn from the 10th Regimental Combat Team, commanded by Colonel William Montgomery. The communication system was almost wholly designed using standard Signal Corps Equipment by a colonel whose first name was "John." He was heavily supported by the Chief Signal Officer, who, when he came to be briefed, replied to the Colonel with "Anything you say, John." Some elements of the Plan were concerned with unit position measurement. ROEC recommended that the areas of the tactical scenarios be overlaid with a 100-yard by 100-yard grid composed of wooden 2 × 4 stakes projecting 5 feet above the ground. Each side of the stake was color-coded with a 5-digit identification number painted on each side. The Corps of Engineers surveyed in, installed, and maintained these stakes. Maintenance, while not great, was irritating because grazing cows tended to push them over. Each field umpire/controller radioed in the position of his attended unit when it moved or engaged the opposing force by estimating his distance and compass bearing from the observed post. He also radioed an estimated range and compass bearing to the target type as well as his estimate of number, exposure, and posture of the target when firing occurred.

FIELD UMPIRE ACTIONS

The space time, response time, and target characteristics were data elements supplied by the field umpires/controllers. The umpire actions were selection and designation of casualties in the unit he monitored from the numbers radioed to him by the Master Control Center located in a tent near the Hearst Mansion. Also he fired simulators for the weapons that did not have smoke and flash simulators. This included mortars, Bazookas, and 106-mm Recoilless Rifles. When he received news of indirect fire on the unit's position he also set off smoke and flash simulators on the point

of impact provided him by Master Control Center. Tape recordings were made of all radio communication among the units.

THE MASTER CONTROL CENTER

According to the CDEC final report on the Umpire Techniques and Procedures Equipment, dated September 1957, the Center was in a tent over an 8-foot by 8-foot horizontal panel with a map representation of the area being played with the numbered 2 × 4 posts identified on the 100 × 100 yard grid. An acetate overlay contained the unit location markings. 2 Senior Plotters with the help of 4 Aggressor and 4 Blue plotters seated to either side of the board kept the positions up to date. Behind, and physically above, the 4 plotters on either side were 3 platoon umpires and one antitank umpire. Well away and operating a second, smaller scale, panel of the playing area was a single indirect fire plotter and one indirect fire umpire. After checking the plotting board, the umpires translated the number of rounds fired, the range to target and target posture into a number (if any) of casualties on the receiving target. This information was transmitted to the appropriate umpire on the other side of the board. This umpire radioed the appropriate field umpire/controller this information. The casualty information was derived from weapons effects data that had been "Monte-Carloed" into a distribution of specific outcomes, using computer time available up and down the West Coast. The platoon umpires had tables of these outcomes for the different direct fire weapons by number of rounds fired, ranges, and target postures. Each time an outcome was used a line was drawn through it with the time and unit identifier noted beside it. When the next similar action was reported the next number of casualties in the list was lined out, etc. Indirect fires and antitank fire received similar treatment. These marked pads were the principal raw data source. They were collected each day.

TANK INSTRUMENTATION AND CONTROL

Tanks were a special problem because of the range and accuracy of their fire. It lay in the fact that the tank umpire/controller could not necessarily know if the tank gun were aimed correctly at a target that the controller in the loader's position might, or might not, see. Broadview Research designed, developed, delivered, and mounted on the tank guns a boresighted, collimated auto head lamp that was turned on by the tank umpire when the tank gunner fired. This light, by adjusting the auto head lamp forward and backward in the collimating tube, could be matched to the .10 to .20 mil dispersion of the tank rounds. The target umpire sighted to see if the light were on the target. If so, he radioed to the Master Control, who told him the extent of casualties to assess. Broadview Research supplied 20 of these lights plus attachment cables for $200 apiece.

GENERAL WYMAN'S VISIT

General Wyman visited CDEC in May when 2 of the 4 phases of testing had been completed. After being briefed, he made the comment, "The whole thing looks too ROCIDy to me." Brigadier General Gibb silenced the numerous protests concerning TRADOC directives and said, "We will change the name to the 'Umpire Techniques and Procedures Experiment (UT&P),' and that CDEC would report only this and not discuss the efficacy of the 'hostile' and 'friendly' unit's organization and tactics."

RESULTS OF THE UT&P EXPERIMENT

The principal effectiveness measure of the UT&P was: The time between one field umpire/controller reporting an action and the time the opposite umpire/controller received the information on casualty assessment. The CDEC report showed that approximately 62% of all these actions in the last 2 of the 4 phases were completed within one minute. It was also concluded that the number of units being so umpired would be limited only by the number of trained umpire/controllers. The radio, radio relay system of tube radios designed and provided by the Signal Corps was deemed adequate. The umpire system tested, thus, had only modest room for improvement. ROEC, probably in violation of General Gibb's orders, explored the range distribution of tank fire between tanks and found it to be essentially the same as that recorded in NW Europe in WWII. Its mean was approximately 670 meters and its median about 500 meters. This, of course, was not reported in the CDEC report. In addition, while no single expression combining enemy casualties, friendly casualties, and time of mission accomplishment was found, there was no need for such a number in evaluating the candidate ROCID and aggressor companies. The phenomenon observed from the Record trials was that for attacking companies (either Red or Blue) in every comparable trial, the attack company that accomplished its mission fastest inflicted the heaviest casualties and suffered the fewest losses. This also could not be reported.

PAUL ERDOS

The major step in the disintegration of the CDEC-ROEC relationship occurred in August or September 1957. I was an invited speaker to a MORS at Stanford, and I was accompanied by the very astute Staff Officer to General Gibb, Colonel Harold Marr. The MORS keynote address was given by the great Hungarian mathematician, Dr. Paul Erdos, who died at 83, this past September. The subject was a very erudite speech on recent advances in Game Theory, one of the most advanced fields of that time. He concluded with: "I see a glimmer, as of the rising sun on a distant horizon, the use of two-sided operational games to predictably measure the outcome of military and corporate operations and strategies." That afternoon I gave my paper on the CDEC approach. When the chair asked for questions or comments, Dr. Erdos strode to the center of the stage, raised his hands high and almost shouted "I was wrong, the sun has already risen high!" Colonel Marr looked at me peculiarly when I got back to my seat. The next day General Gibb called me in his office and made the following points, which I believed then and believe now were absolutely sincere:

ONE: He felt deceived because I had not told him that the two-sided operational game had never before been tried. (Note that I had used this term, from the first, to describe the test model).

TWO: He believed that Good Science was the application of established and thoroughly proven methods.

THREE: He had expected that use of the Scientific Method and Objective Science would reduce the effort required to develop Organizations and Procedures. Rather, he had to drive his staff and troops harder than he did when he was with the 1st Infantry Division in WWII.

FOUR: He felt that he could no longer place complete trust in the ROEC scientific support, and would seek outside expert help in future CDEC work.

REPLACEMENT ON THE 100 × 100-YARD STAKES

The stakes were replaced, over my warning, with 500-yard by 500-yard 20 foot steel posts carrying a large box on top with the identifying markers. In the Mobility experiment it was frequently found that the boxes were not found or could not be seen by the field umpire/controllers. The space time position measurement now had big errors due to search problems (Shades of Project STALK!!) and trees whose branches spread at 10 to 15 feet above the ground. Moreover, when the new stakes were observed the average umpire/controller estimation of distance from his true position was increased from about 10 yards to about 50 yards. Once again it was found that mean range estimation error is about 20% of true range! Remember the Corps of Engineers did not like the wooden stakes. In addition, based on a proposal by IBM-San Jose, an IBM/620 (1620) computer was installed at the Master Control Center and a computer controlled vertical back lighted panel replaced the Master Control board. This provided a good view for visiting dignitaries but was of diminished help to the controllers. Most importantly, the 1620 was used to solve the Monte Carlo selection of casualties from weapons effects functions. The resulting queuing increased the measure of time from field umpire input to message receipt by the umpire of the opposing force from 1 minute to 6 minutes for 62% of the casualty assessments. Position recordings from the 500 × 500 yard posts also queued at the computer. Something had to be done.

TRILATERATION

In the face of my remonstrations, CDEC took proposals from such contractors as Cubic Corporation for optical or electronic trilateration schemes. I proposed using the British Bendix low-frequency hyperbolic grid scheme being used in Portsmouth Harbor with the expectation it could be used in hilly, tree-covered terrain. It was rejected because it was British. (More shades of Project STALK where the Centurian tank was not allowed by DA because it was British). Clearly, of course, trilateration requires movement of the forces being tested into open terrain.

DR. IAN TERVETT

In the Fall of 1977, Dr. Ian Tervett replaced Dr. Frank Brooks as Director of ROEC. Dr. Fred Henriques of TOI did this because of Frank's health problems. Dr. Tervett had recently left the U.S. Army Chemical Corps as a Civil Servant, where his major research work had been on Chemical Defoliants. He had some experience in testing them. He strongly felt that General Gibb needed a Civil Service Chief Scientist—a position he took shortly after the demise of the ROEC contract in June 1958.

SCOUTING AND HELICOPTER EXPERIMENTS

In the last year of its contract, ROEC designed and supported several tests. Among these, was the Scouting Experiment, where it was found that the number of hostile detections by U.S. Army scouting units had a very high correlation with the number of scouting observers regardless of their

mode of transportation including: Jeep, armored vehicle, helicopter or foot; or the scouting tactics used. Line of sight was a necessary, but insufficient condition for target detection. (More shades of Project STALK!) A helicopter-borne scout was recorded as acquiring twice as many detections, if the detections by the pilot were not taken into account. I have no copy of this CDEC report if ever one was made. I recall serious criticism because the PPS-4 was not used. This was because it could not function after Jeep transportation to the test site. It was a very early model. A Helicopter Experiment was run using gun cameras and helicopter vulnerability data from BRL for various anti-aircraft and other direct fire weapons such as rifles. The controversial finding was the low level helicopter flights in variable terrain would expose them to being hit frequently. In 1958, the conventional wisdom was that could not happen because of our experience in Korea, where the CAA did not fire on our medical helicopters removing Allied and Chinese casualties from the battlefield. This was not reported by CDEC. Rather the Operations Research Office was brought in to design and execute essentially the same test using Photo-Theodolites. ORO's Draft report on its test came out about 4–6 years later, as I recall. In any case, its evidence was the same.

RELOCATION

Although I was proffered a job by SRI, as was the rest of the technical staff, and one in Boston by TOI, I moved to Virginia in 1958 to work on a Combat Surveillance Contract of Connell Aeronautical Laboratories with the Signal Corps Combat Surveillance Agency. I went to work on the concept definition and testing of the SD-2 Surveillance Drone.

THANK YOU!!

Example of a 4 × 4 Graeco-Latin Square

CAR

DRIVER	1	2	3	4
I	Aα	Bβ	Cγ	Dδ
II	Bδ	Aγ	Dβ	Cα
III	Cβ	Dα	Aδ	Bγ
IV	Dγ	Cδ	Bα	Aβ

Additives: A, B, C, D
Days: α, β, γ, δ

FROM FIELD EXPERIMENTATION TO SIMULATION: THE FORTY YEAR QUEST TO UNDERSTAND COMPLEX SYSTEMS

By
Henry C. Alberts, Professor of Acquisition Management
Defense Systems Management College, Fort Belvoir, Virginia 22060

ABSTRACT

From its founding in 1956, the experimental facilities established by the Army, first at Fort Ord and then at Fort Hunter-Liggett California as the Combat Development Experimentation Center (CDEC) have provided a unique laboratory to explore the behavior of complex systems. At first, with the most rudimentary information providing equipment, and later with more modern devices, the events and relationships among elements of fighting forces were played out on the field in disciplined activities which helped provide crucial insight: (1) for our armed forces in combat; (2) for those who devise operational tactics; and, (3) for those who plan and design new combat equipment.

This paper examines the years between 1962 and 1981 from the point of view of the instrumentation capabilities used to provide data upon which analyses were based and traces the increasing sophistication of data collection and management devices throughout the period.

A PERSONAL VIGNETTE

In 1956, the Army's Chief of Ordnance contracted with the Pennsylvania State University (PSU) to form a team to study then available U.S. capability to defend the United States against threats posed by Intercontinental Ballistic Missiles (ICBMs). I was a member of that study team. My qualifications included: (1) research and experimental work in supersonic flow phenomena which I had done for the Army at the Ballistic Research Laboratory during the years 1949 through 1953; (2) service as coordinator for Geophysical Research and Development involved with the U.S. Air Force's Guided Missile activities from 1953 to 1956; and (3) experience as Head of Operations Research for AVCO Corporation's Advanced Research and Development organization, Air Force contractor for design, fabrication, and test of the ATLAS ICBM re-entry body.

It was in connection with this latter work that I had performed a study of the vulnerability of ICBM vehicles to existing anti-missile systems. The PSU principal investigator, Dr. Harold Hipsch, Chairman of the Aeronautical Engineering Department, was extremely interested in my experiences in design and construction of re-entry vehicles. He wanted to develop a "time-line" of events which could be used to estimate which of the multiple potential defense configurations would likely be most effective. I had done this for the ATLAS missile, making estimates of elapsed times between significant events. I had also examined operational sequences of missile preparation, launch, flight, re-entry, and impact. I had attempted to find "hard data" relevant to each phase of missile operations. But although there were attempts to collect measurements during the normal course of development activity, there were no organized, consistent programs which had as their objective the disciplined design of reproducible sequences of events.

Consequently, the estimates of duration of time-line events were derived from theoretical considerations, which we later found to have little resemblance to the actual operating context.

One of the PSU team members was a Lieutenant Colonel of Infantry. He spent his study team time reminding us that we were working to improve the capability of "real soldiers", and that the theory we were developing would need to be applied to the real world of actual field maneuver. But we all recognized that the capability to do that was limited by the existing, available field facilities. One day, however, he informed us of the establishment of a new proving ground complex in California which would explore the relationships of individual troops engaged in simulated operational maneuvers.

I heard very little about the result of initial activities at CDEC until 1959, when I began to work with Dr. William C. Pettijohn who was then at Johns Hopkins University's Operations Research Office. One of AVCO's products was a shell designed for the Army's new M-79 grenade launcher, and there had been difficulty in maintaining both the required CEP and the round to round dispersion when firing production ammunition. As Head of Operations Research, I was asked to look at the problems and see how to solve them. Dr. Pettijohn arranged with me to perform a field measurement program to examine the characteristics of the M-79 weapon with specific emphasis on how well soldiers using it could aim their fire. When he completed the work, we found that the sighting and aiming errors were so large that the requirements for tight CEP and low shell round-to-round dispersion would severely limit weapon effectiveness: The aiming errors were between 20% and 25% of range to target! In the process of performing the experimental work, Dr. Pettijohn became very interested in the concept of CDEC and how CDEC's type of activity might materially improve Army combat capability. He moved to CDEC shortly thereafter, working for Stanford Research Institute in the Fort Ord Research Office.

In 1960, I joined National Company, Incorporated, which had been engaged in developing state of the art communication equipment, and super-accurate timing devices. One of the products was the first "atomic clock". That particular device was based on a Rubidium gas standard and used the quantum energy available from excitation of the Rubidium atoms to maintain a digital counter accurate to tens of milliseconds per year. We used the clock to perform a test of the capability to synchronize time across great distances; and incidentally, by using a B-36 platform in continuous flight, we were able to check on Einstein's Twin Paradox. - the prediction that a rapidly moving platform experiences times passage slower than one which is stationary. In 1961, I visited Stanford Research Institute to brief the staff on the results of what we had called "Operation Time Tack." Dr. Pettijohn and Scroggie Wiley provided *quid* for *the pro quo* by talking at great length about the difficulties they were experiencing in collecting time related data at CDEC. Believing that I could contribute to that problem's solution, I joined the SRI Research Office in October, 1962 to work on improving the capability to instrument the field activities and permit collecting integrated, time-sequenced position location and event information.

FROM FIELD EXPERIMENTATION TO SIMULATION

STEPS IN THE PROCESS OF MEASURING REALITY

The process of designing an instrumentation system for CDEC experimentation began with devising a plan which would:

1. Measure the: (a) experimental time line on which all events could be placed; (b) position all of the men and equipment on the field (which we grouped under the heading of "players"); (c) events which took place at every location in the field; (d) results of all engagements among the players

2. Insofar as possible, provide a degree of field realism to try to make players respond as if they were in real combat action.

3. Provide the capability to capture, classify, evaluate, and display the collected information in real time to the large numbers of individuals involved with directing, monitoring, and analyzing the field activities in progress; and reconstructing the action repeatedly so it could be studied in detail.

I had thought that the fundamental issue of providing for a synchronous experimental-test time line upon which to place each event taking place in the field would yield easily. After all, the Naval Observatory routinely broadcast the U.S. standard timing signals over WWV. But that hope soon faded. There were propagation anomalies which made Fort Hunter-Liggett unsuitable for standard kinds of then available broadcast systems used to send time across space. In the end, we were forced to provide and broadcast a timing signal from the experimental area to experimental participants - and even that specially designed system could not provide timing signals throughout the experimental terrain. Nor could other kinds of radio signals be reliably sent from those areas to a control center location.

For similar reasons, it was infeasable to use standard navigational systems such as LORAN or TACAN to provide position measurements at all player locations. We were required to construct our own triangulation mechanism and to devise specific kinds of player modules for field use.

Additionally, using the newly developed position-location equipment to transmit events which took place at the players' locations turned out to require more bandwidth than was available on an already restricted transmission frequency set.

The problem of marking engagement pairs, and assessing the results was also challenging. Here the difficulty was in determining whether an engagement could even have occurred: Did line of sight exist between the two players? If an indirect fire engagement was in progress, did the settings of the weapons and the positions of potential targets enable fire coverage in the particular parts of the terrain involved? Lacking the motivation of live fire, were actions taken by target players representative of their responses in actual combat?

Displays, too, were unable to assimilate and process the large amounts of information which would be taken in the field when all of the instrumentation were working and reporting at the data repetition rates we thought we needed to ensure performance of analyses of the desired accuracy.

At the time, I characterized the problems of devising a useful, real-time field data collection system as "trying to do 21st Century science under Medieval field conditions!"

By mid-1963, progress had been made in all technical areas required to provide the basis for instrumentation development. And then, in the midst of it all, there arose a debate between The Ballistic Research Laboratories at Aberdeen, and those who were developing vertical envelopment tactics in Vietnam related to the survivability of large numbers of helicopters operating in that environment. We were tasked to develop and execute an experimental plan to obtain data on the effectiveness of ground live-fire against helicopters. One part of the program required us to develop live fire targets that looked and maneuvered like UH-1B aircraft, and that could report the event of their having come under fire. In addition, if the targets were hit, they would also report the location of projectile entry and exit so that the probable aim and firing information could be captured.

The component instrumentation for the test program was developed and in place in less than 9 months. Drone helicopter targets were constructed using reconditioned OH-13 units (purchased from oil rig operators) fitted with a hit sensitive skin which made them resemble a scaled down UH-1B, and carrying an array of microphones which permitted acoustic measurements of the shock waves emanating from projectiles which approached the envelope of sensitivity around the target. The firing itself took place at Fort Bliss Texas in late 1964 and 1965. Use of Fort Bliss allowed us to use the position location and timing instrumentation in place on the Dona Ana Range. We learned a lot from this program and became involved in a debate with BRL related to the process of aiming and firing multiple shot and automatic weapons. We predicted higher survival rates for vertical envelopment tactics in Vietnam than had been predicted (and accepted as likely) by them and others. When our estimates were confirmed in action, we felt we had made a real contribution to our fighting forces.

From the perspective of instrumentation, digital computers were in their primary stages of development at that time. Only in 1960 did digital process become the preferred methodology to perform complex computations. Prior to that time, analog computers were used to represent systems and to determine results of varying any of the many parameters involved in their performance. Only eight years had elapsed since Dr. William Shockley had demonstrated the capability of doped silicon wafers to act as amplifying devices. The entire concept of digital communications as a replacement for the standard analog transmission theory and communications construction methodology was still some time in the future. In many ways, attempts to achieve the objective of providing measurements which would allow us to understand

very complex systems resulted in our having to advance the state of the art in sensing, communications, display, and mathematical analysis. It didn't seem as though we were on the cutting edge of technical capability - although when we would talk with others who were attempting similar things, we found out we were. To us, it seemed as if the ancient Chinese curse had come to pass: We were "living in interesting times!!"

Although there was significant progress toward developing a capability to measure field occurrences and perform analysis of them, we found that we were doing many of the same experiments over and over again. I asked the Research Board to consider the possibility of constructing a series of experimental building blocks: exercises which would be performed under broad sets of conditions and then used as "ground truth" for those elements for ever after. I had the idea that Omar, the tent maker was correct when he said in the Rubiyat: "The moving finger writes, and having writ moves on: nor all thy piety and wit shall lure it back to cancel half a line, nor all thy tears wash out a word of it!" How naive I was. I had no understanding then that when dealing with large complex systems, there is only limited, if any, reproducibility over time and space.

As we continued to devise instrumentation to capture the operational world as represented in the field at CDEC, we began to sense other problems made visible by the considerable improvement we had made in measuring sequences of events on a consistent time line. We saw that there was considerable variance in performance of set-piece tasks depending upon precedent and antecedent tasks. We tended to minimize these variances and declare "experimental error" as the cause. I failed then to grasp the full meaning of what I saw, Later, I would be able to place it all into perspective and draw insight from the experience: I learned that outcomes of complex events are highly dependent on the task sequences and the life experiences of individuals involved in their performance. This would emerge more clearly in work on Small Independent Action Forces discussed below.

In 1966, I left the Research Office to work with a former CDEC Commandant, BG Charles J. Girard who was assigned to Headquarters, Seventh Army in Germany. SRI provided a team of analysts to consider the problems involved in using information developed in yearly field exercises. When instrumenting CDEC, I could foresee a day when there would be a plethora of data; a time when an individual human being could not comprehend everything the instrumentation would tell them. Psychologists call this kind of problem, "Cognitive Overload" and it is a common occurrence in today's world. In Germany I learned what too much data, both organized and unorganized, could do to human understanding. We had just about completed the work in Germany when Braddock, Dunn, and McDonald (BDM) assumed responsibility for CDEC support.

I worked on a number of problems at SRI Menlo Park before moving to Sweden in October of 1966. Once there, I worked on private sector problems. In the process of serving private sector needs, I learned a great deal about the difficulty in obtaining, interpreting, and presenting "data"

so that it provides "Information". As a result, when I returned to the United States in 1968, I was able to understand that there were extensive commonalties between the commercial and military worlds: in both it was difficult to deal with large information flows generated from real time observations of complex systems.

One practical illustration of how "field experimentation" and "field exercise" can turn into "simulation" has its roots in the CDEC experience. It concerns a large data base building program based on data collected about Small Independent Action Forces (SIAF) operating independently of Battalion and Division control in Vietnam. Small units had been seen to be more successful in detecting and reporting enemy activities, engaging when necessary, all while keeping casualty counts below units operating interior to large force elements. The field experience of Army Long Range Reconnaissance Patrols (LRRPs), Navy SEALS, and Marine Reconnaissance Units had shown that their operating tactics provided an effective means of combatting both the North Vietnamese Army, and Viet-Cong forces. The people at ARPA wanted to understand exactly how small unit force actions differed from larger scale fighting and to use that understanding to develop better force deployment and action tactics. Dr. Pettijohn had already joined ARPA's support contractor team, and I joined him there in 1970. Together, we spent three years building a data base which could describe quite accurately the way small units operated in Vietnam. Building the data base required 5 step process:

1. interviews were conducted with members of the U.S. SIAF units immediately after they had returned from patrols. Each patrol member was asked to reconstruct the entire patrol experience from his own point of view. The applicable terrain maps were laid out and questions were asked about: (a) the Operational Order; (b) the actual insertion; (c) how the patrol proceeded on the ground; (d) how fast it went across the terrain; (e) how many enemy detections were made which did not result in engagement, and the circumstances under which they occurred; (f) the fire fights (if any) which resulted from enemy detection of friendly forces either prior to, or concurrent with detection by friendly forces and the expenditure rates of ammunition during those fire-fights; (g) the external support required (h) the withdrawal; and, (i) the perception of patrol results. Additional interviews were conducted with small units made up of foreign troops who were operating independently of larger units to gain comparable understanding of how they functioned during their patrols.

2. Pictures of representative terrain were shown to patrol members and the data they had provided in their interviews was linked to the terrain type over which the patrol proceeded during each time increment. Patrol members were asked to explain the reasons why they would select a movement rate, what dictated their positions during both movement and at rest, their estimate of the density of enemy troops on the terrain, and the way in which the enemy dispersed and moved over the ground during the patrol period.

3. A statistical analysis was performed to derive relationships among: (a) terrain types; (b) terrain movement rates' (c) perceived enemy distributions over terrain; (d) detection occurrences and

probability for both friendly and enemy forces; (e) engagement probability given detection; and (f) outcomes of engagements (including rates of ammunition expenditure, numbers and types of casualties).

4. Seasoned Vietnam veteran troops were used in a field experiment in Hawaii National Forest - an area on the island of Hawaii which resembled an area of Vietnam about which we had gathered considerable information. "Enemy" troops were dispersed on the terrain in tactical positions corresponding to those used by both North Vietnamese and Viet-Cong forces; and these troops moved on the terrain tactically as those enemy forces would have done. 24 small independent action force patrols were asked to perform search and reconnaissance missions over the terrain, and their movements and all other activities monitored with considerable accuracy. Each patrol was of five to eight days duration. At first light each morning, the data collected from the previous day's activity was flown to Honolulu and processed on a CDC 6400 computer. Results were returned to Hawaii as soon as they were obtained, usually prior to 3:00 P.M. of the same day. Activities for the next day were determined based on the totality of data processed up to that time.

5. A computer assisted game was developed. Complete with terrain film clips, operational orders, simulated enemy troop distributions and movements. The intent was to provide a simulation of the experience captured within the data base. The simulation was applied to twelve experienced Vietnam combat patrols at the Special Forces Training School.. As the simulations ran, data was taken about troop responses.

6. The data from Vietnam, Hawaii, and the Special Forces School were compared to see if each data set belonged within the same data universe. When that had been shown, we had considerable confidence that we would be able to test new tactics in simulated Vietnam conditions without deploying large numbers of troops in field experiments specifically for that purpose.

When the SIAF work ended, I turned to other kinds of data collection and analysis work. With the exception of planning two more field experiments for the Marine Corps Development and Education Center, (MCDEC) during the period from 1974 through 1980, I was absent from the military field experimentation milieu.

In October 1980, Scroggie Wiley called me and asked if I would be interested in returning to CDEC in my old role as Director of Instrumentation. I was delighted to do so. I had so enjoyed my time at CDEC that I was happy with the idea of reliving my youth. When I returned in November 1980, I saw that the things we had pressed as advances to the state of the art in field data collection in the early and mid-1960's had been completed and were functioning. I looked forward to defining the next generation of data collection, processing, and analytical devices and to putting them into the field to achieve distributed systems: a kind of internet concept for experimentation where simulations resident externally to CDEC would be incorporated within

CDEC's field experiments and results obtained through exercise of simulations transmitted to CDEC for use within ongoing experimentation. Even at that time, it seemed clear that the cost of the sort of activities in which CDEC has been historically engaged was rapidly becoming prohibitive. And it was also clear that some ideas for improvements to existing CDEC instrumentation equipment had been institutionalized within the experimental community to the extent that progress in making great change to historical directions would be difficult to achieve.

About the middle of July 1981. it also became clear that my family was firmly anchored to the Eastern half of the united States. All of my children and grandchildren were there. With the desire to maintain close family ties uppermost in my mind, I returned to Virginia in November 1981 and, in 1983 joined the Defense Systems Management College as Professor of Engineering Management in the hope of helping students cope with the very complicated business of design, development, test, production, and support of military weapon systems.

AND WHAT ABOUT TOMORROW?

My grand-children ask me about historical things. They say I am "Living History". It is as hard for them to understand the world in which I grew up as it is for me to understand what life was like when my parents were young. As we have all come to know, perceptions are fact; truth is transient; and the future is a guess! Notwithstanding all of that, I would like to make some guesses to this particular audience about the effect of technology's relentless advance and how we gain understanding of complex activities (of which warfare is certainly one of the most complex and intense of human activities).

Marshall McLuhan [1] provided us with the insight into how the Russian empire would fall. He projected a "global village" in which information could not be controlled and where the power to see what was happening, as it was happening, would inexorably shape world events. There are few who would deny that continuous presentation of scenes of war and death on the evening news accompanying dinner was a major force in shaping the policy which led to disengagement in Vietnam. And my Russian friend (whom I am now free to know and work with) tells me that the USSR was doomed the first time Russian citizens saw the Western way of life on television and found that their government had been consistently lying to them. The visual evidence of "the way things are" transcends even long-held opinions about "the way I have been told things are."

Similarly, it has been writers of fiction who have presented an "envisioned" future "outside of the boundaries" of today's realities. Perhaps they are most likely to be closer to the things which lie in our future. Today, aircraft simulators present pilots with extremely realistic presentations of flight. Technology has permitted movements from the first simulators used for pilot training in the World War II era, through the more sophisticated devices created by the Naval Special Devices Center in the late 1940's and early 1950's to the combat training devices pilots use to prepare themselves for flying the high performance supersonic aircraft of today.

Just as Jules Verne [2] predicted a nuclear submarine many decades before nuclear energy was even conceived of by science, and the many creators of the character of Buck Rogers were correct about man's voyages into space, Gene Roddenberry, the creator of the original "Star Trek" might have presented a vision of how warfare may be conducted in the future. In an episode of "Star Trek", the crew of the Enterprise finds itself in orbit around a planet which is at war with a neighboring planet. In this episode, the war is fought on computers: moves are programmed into the computers of both combatants, casualties are computed, and each government responds by ordering the proper numbers of people to be killed. While we might not carry realism as far as that, we can now create very realistic displays of combat which present players with the illusion of direct personal involvement in a field action. We can bring together in a virtual network, many individual weapon simulators and devices which add to the verisimilitude of tactical situations. Communications and display have advanced to the point where Roddenberry's vision can be implemented!

But there is a further possibility in the immediate future that can do even better. At the Vancouver World's Fair of 1984, the story of the development of British Columbia and its natural resources was presented in the B.C. Pavilion. One of the vignettes (for which there was always an extended queue) had to do with Indian tribes and their lives before the settlers came. I sat in my seat facing a large stage shielded by curtains. The curtains parted and an Indian village was revealed complete with a complement of teepees, a lake, and groups of people who inhabited that area. An old Indian (a tribal elder) came on-stage and told the story of the village and life there during the days before settlement. At the end of his story, all of the other people walked off the stage, and only he was left. When he had finished his story, a canoe came floating in from the rear of the stage and came to a stop in front of him. As he spoke about the disappearance of the Indian way of life and of the people who fulfilled the expectations of that life, he slowly climbed an invisible ladder, sat down inside the canoe, and floated off at the rear of the stage as the village and the scenery ALSO disappeared from view. It was only after he had vanished and the empty stage revealed, that the audience became aware it had been watching a holographic spectacle so realistic that it had been mistaken for a live performance! We had seen "virtual reality" made possible holographically to a large audience who were so immersed in the illusion that they felt themselves part of the spectacle.

Imagine what can be done today with computer generated virtual reality! Together, linked interactive communication networks using distributed simulations blended together in computer-driven three-dimensional (or even holographic) displays can make it possible to immerse individuals in simulated battle so realistically that their responses can be used to provide highly reliable indications of an actual battle outcome. The capability to perform parallel computer processing at high speeds can provide seamless simulations of sufficient dimensionality to drive virtual reality displays which place players inside of "a world that could be." Field exercises or experiments would provide confirmation of the simulation outcome rather than basic information

from which a simulation could be constructed. Realism can be achieved through creation of battle reality sufficiently well to generate believable responses appropriate to the situations created.

Under such circumstances, the purposes for which CDEC was established 40 years ago can be fulfilled in many different localities; and the need to set aside large numbers of troops dedicated to performing experimentation missions becomes much less necessary. While there will likely need to be some specially-instrumented locations specific to the purpose of test and evaluation of real equipment by real soldiers on a real terrain, the majority of that work will most likely be possible at control-room kinds of locations distributed throughout the United States.

In short: We will have achieved the capability to perform continuous, controlled experimentation in an orderly exploration of the effects of changing equipment and tactics on battle outcomes!

As for the CDEC I knew and loved, perhaps an old Latin phrase may be appropriate: ***"Sic transit gloria mundi!"*** For those of you who are not Latin scholars, it means: "Thus passes the glory of the world"

Henry C. Alberts
Bethany Beach, Delaware and Fort Belvoir Virginia
October, 1996

REFERENCES

[1] *The Medium Is The Message*; Marshall McLuhan

[2] *Voyage To The Bottom Of The Sea*; Jules Verne

[3 *Star Trek Episode*; Created by Gene Roddenberry

Modeling Loss Exchange Ratios as Inverse Gaussian Variates: Implications

D. H. Olwell
Department Of Mathematical Sciences
U. S. Military Academy, West Point, New York 10996 [*] [†]

December 10, 1996

Abstract

This report outlines methods for estimating and comparing the Loss Exchange Ratio (LER) output of computer combat simulations, and develops methods to establish a priori the number of simulation runs required to detect a change in the parameters of the simulation of a given size.

The Loss Exchange Ratio (LER) is a widely used and widely accepted summary statistic for a simulation run involving force-on-force combat models. The LER is surprisingly variable – multiple runs of the same scenario produce a large range of LER.

We assert here that these loss exchange ratios are skew stochastic random variables, and that they are well modeled by the inverse gaussian (IG) distribution. We discuss technical reasons for preferring the inverse gaussian for models over other distributions, particularly the log-normal distribution.

Adopting this IG stochastic model allows us to develop explicit statistical methods for estimating the parameters of this distribution, using its known sampling distributions. We also inherit the precise statistical tests for hypothesis testing. Finally, we are able to determine a priori the number of simulation runs necessary to detect a change in the distribution of a given size. This is a particularly valuable ability, given the increased reliance of the Army on these simulation models to make procurement and doctrinal decisions. We discuss how these simulations test fit into the larger scheme of procurement and doctrine decisions

We illustrate with data sets from both the JANUS and CASTFOREM simulations. In particular, we find that the use of the IG model allows us to make more powerful conclusions about the data.

We conclude that the IG is a good model for describing the variability of LER with useful estimation and testing properties, and recommend its adoption. We sketch several other promising areas for research which follow from the adoption of this model.

Key Words: Sample size, loss exchange ratios, inverse gaussian random variables, JANUS, CASTFOREM, simulation, design of experiments

[*]This research was supported in part by the Army Research Laboratories and the Mathematical Sciences Center of Excellence, USMA.

[†]Approved for public release; distribution unlimited.

Introduction

Consider two systems which are being considered for acquisition. How does one tell if they are worth the cost of acquiring them, or what their benefits are? The question is particularly difficult if the systems are from different battlefield operating systems, say an air defense weapon system and a communications system.

One strategy for comparing these systems is to model their characteristics, and add them to an existing "base case" force model. For example, we may have a force model which represents a battalion task force. We adjust the model to reflect the addition of new, competing weapons systems. These new force models are used in a suite of scenarios which are executed in a combat simulation, say JANUS or CASTFOREM. The results of the simulation with the new force packages are compared to each other and to the base case. Inferences are drawn about their relative merits. These merits, together with the costs of the systems, can form the basis for rational choices using a cost–benefit analysis.

Such comparisons are not limited to system acquisition: doctrine and force structures can also be modeled and compared using this simulation approach.

A related problem asks, what are the specifications which should be required for a new system? One approach is to construct a model which allows varying capability in the new weapon system, and to simulate at various levels of this capability. One then chooses a response from the simulation, and constructs a model of the response as a function of the level of the capability. It is possible to construct response surface models which examine the effects of making multiple changes to the base force model simultaneously. These models help decide how much and which capability to buy. They also allow exploration of the interactions between capabilities and the identification of any resulting synergies.

This methodology requires us to select the outputs of these simulations for comparison. It then requires a statistically valid means of modeling, estimating, and comparing these responses.

One of the conventional summary statistics for a combat simulation run is the loss exchange ratio (LER) , which is the ratio of enemy losses to friendly losses. While this statistic suffers from all the difficulties associated with summarizing a very complex battle with one number, it has found wide acceptance in the operations research community.

We will use the LER as the response variable for the purposes of this discussion. We note that the methodology is general, and can be applied to other skewed, non-negative measures of effectiveness.

This paper has the following structure. In section 2, we discuss loss exchange variables and possible models, adopting the inverse gaussian model. In section 3, we discuss estimation of LER parameters using the inverse gaussian model. In section 4, we discuss hypothesis testing. In section 5, we discuss sequential testing methods of LER. Next, in section 6, we examine the power of these tests, and propose a simulation method for determining the appropriate number of runs for a simulation. We then look at two different sets of simulation results in section 7. We close with conclusions and recommendations. A primer on the inverse gaussian distribution is available from the author, and is omitted here in the interests of conserving space.

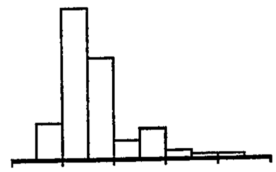


Figure 1: Histogram for loss exchange ratios for 80 simulations of a scenario simulated using CASTFOREM. Data provided by TRAC–WSMR.

Loss Exchange Ratios

The loss exchange ratio is a widely used summary statistic for combat models. It has theoretical underpinnings in the work of Frederick Lanchester, and his deterministic differential equation models of combat.

It is well known that the output of a computer simulation package such as JANUS or CASTFOREM is variable. The exact same scenario can be simulated repeatedly on these models, and different – sometimes strikingly different – outcomes may result. For example, the boxplot in Figure 1 shows the LERs of 80 runs of the same scenario on the same computer using the same simulation package, CASTFOREM. The maximum LER was 4.5, while the minimum was 0.69. The median was 1.5, while the mean was 1.69788.

The variability and skewness in these data argue strongly against using the single summary statistic, average LER. The data needs to be described not only with a measure of location, but also with measures of its dispersion and shape. For appropriate statistical description and analysis, we require a statistical model. Lacking such a model, we can not compare the outputs of the competing simulations: we can not determine if the difference in response is due merely to chance.

Models

There are several possible models for modeling non-negative skew data. The log-normal, gamma, weibull, and inverse gaussian distributions immediately suggest themselves.

We desire our model to have several properties. First, the model must fit the data well. Second, the distribution of the maximum likelihood estimators (MLEs) for the model parameters should be known, and tractable. As a minimum, we should be able to find the MLEs without resorting to numerical methods. Third, the theory of estimation and testing for the model should be well developed. Fourth, the parameters of the model should be easily interpretable.

We exclude the gamma and the weibull distributions for failing to have the second property. The MLEs for these distributions can not be found explicitly, and require numerical approximation. The distribution for the MLEs is not tractable.

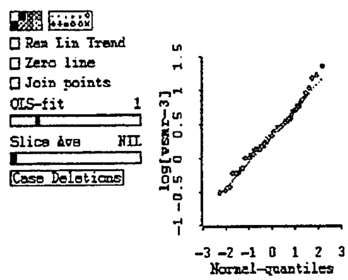


Figure 2: A "QQ" plot of the logarithm of the WSMR data against normal quantiles.

The log-normal is a possible model. The distribution of the MLE's is known, and parallels the standard normal distributions. However, there is a real practical difficulty which arises from the logarithmic transformation of the data necessary to conduct statistical testing. Statements about the mean of the transformed variable are not statements about the mean of the original variable, but rather the median of the original variable. The mean of the original variable is a function of both the mean and variance of the transformed variable. A direct test for equality of means of the original variable is awkward at best. Similarly, statements about the variance of the original variable are complicated by the fact that it is a function of both the mean and variance of the transformed variables.

We prefer a model which fits well and does not require transformation, so that the parameters are immediately useful. As we discuss in the next section, we choose the inverse gaussian distribution.

Why Inverse Gaussian?

The inverse gaussian distribution is a positive skewed distribution with two parameters, μ and λ: μ is the mean of the distribution, and λ is a shape parameter. The MLEs are known and the distributions of the MLEs involve only the inverse gaussian distribution and the Chi–squared distribution. Statistical tests for equality of μ and λ involve only the t–distributions and the F–distributions.

The inverse gaussian distribution fits the data sets we display in this report at least as well as the log–normal. The difference between the two is in the behavior of the left tail, where the log-normal tends to underestimate the quantiles.

For example, Figure 2 is a "QQ" plot of the data set from Figure 1 against the log–normal distribution. Notice that the data is more heavy tailed to the right than the normal quantiles would suggest. Similarly, Figure 3 shows that the histogram for the transformed data is still skewed to the right. Figure 4 shows the density for the inverse gaussian distribution with MLEs, and the model fits the tails better.

The graphical evidence in Figures 2, 3, and 4 is supported by more formal goodness of fit testing using the Wilks-Shapiro statistic.

We make the assumption for the balance of this paper that the LER data is well modeled by the inverse gaussian distribution, with parameters given by the maximum likelihood estimates.

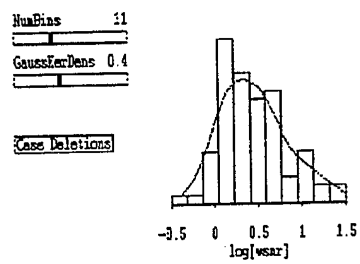


Figure 3: A histogram of the logarithm of the WSMR base data set. A non-parametric smooth has been applied to the data. Notice that the data is still skew to the right, suggesting that the log-normal model is inappropriate.

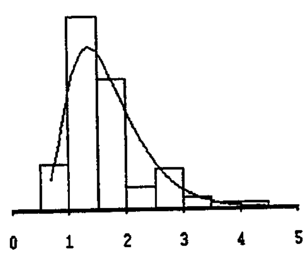


Figure 4: Histogram of the WSMR base data with best fitting IG density.

Estimation

The MLE estimate of the mean of the IG distribution is the sample average and its distribution is $\overline{X} \sim IG(\mu, n\lambda)$. This allows us to construct confidence intervals for the mean of the LER. These confidence intervals are more accurate than ones based on the asymptotic application of the law of large numbers, because the data is more heavy tailed than the normal distribution. Application of the standard $\overline{x} \pm k\hat{\sigma}$ results in an unnecessarily large confidence interval for the mean.

The shape parameter λ has MLE given by

$$\frac{1}{\hat{\lambda}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{X_i} - \frac{1}{\overline{X}}\right) = \frac{1}{n}V \tag{1}$$

This estimator is a function of the sufficient statistic V.

The distribution of $V \sim \frac{1}{\lambda}\chi^2_{n-1}$. This allows confidence intervals to be constructed for λ based on the χ^2 distribution.

For further details, the reader is referred to the primer in the appendix.

The key point is that the distributions of these MLEs involve only the IG and the χ^2 distributions: they are very tractable. The actual estimates are easily computed.

Estimating the shape parameter seems to be particularly noteworthy, as the skewness and variance of the LER are not routinely reported. The variance of the $IG(\mu, \lambda)$ distribution is given by μ^3/λ, so as the shape parameter increases, the variability decreases.

Closed form expressions for confidence intervals for μ and λ are available in Chhikara and Folks [1989], and again are based on quantiles of standard distributions.

These confidence intervals are narrower than ones based on the asymptotic normal distribution. For example, consider the WSMR base data. A 95% confidence interval, based on a standard normal distribution approximation for the mean which follows from the strong law of large numbers, we obtain

$$\mu \in (1.53895, 1.8568) = (\mu \pm 1.96\sigma/\sqrt{n}) \tag{2}$$

Using the IG model and the formulas given in Section 9, we obtain a tighter confidence interval for μ, which also recognizes the skew nature of the data:

$$\mu \in (1.56257, 1.85883) \tag{3}$$

As a result, we have a more precise estimate of the mean, given the available data.

We obtain similar confidence intervals for the λ parameter.

We mention in passing that we are averse to confidence intervals for the mean, preferring instead to assume a Bayesian model with a non-informative prior distribution, which results in probability intervals for the mean. Such Bayesian methods are also outlined in Section 9.

Hypothesis Testing

The uniform most powerful unbiased tests for the equality of two inverse gaussian population means are known. We consider here the case where neither the mean nor shape parameter is known. References for the other cases are in Section 9.

The rejection region is a function of the sufficient statistics for each sample, $\overline{X}$ and V, and the critical points are given by the t distribution. Details are given in the primer in the appendix, Section ??.

The uniform most powerful test for the equality of the shape parameters is a function of the sufficient statistics V for each sample, and follows the F distribution. Again, details are in the primer in Section ??.

These tests allow us to test if the means and shapes of two samples are statistically equivalent. In the context of our problem of comparing the output of two combat simulations, they allow us to test the hypothesis that the outputs came from identical processes.

Moreover, since these tests are based on well fit distributions, they are more powerful than using asymptotically based tests. We see in the examples where these tests allow us to show statistically different results, where the asymptotic methods do not.

The result is that we can make more powerful inference based on the simulations we do run, which saves us computational expense and results in more efficient use of the simulations we do run. For large simulations, this can result in significant economies.

Significance tests also exist for one sided and two sided tests for the mean with λ both known and unknown. Significance tests also are known for the equality of λ with the mean both known and unknown. Additionally, there are two sample versions of the above tests. These cover the usual possibilities, involve only the IG, χ^2, t, and F distributions, and allow simple implementation of exact tests. These tests are outlined in Chhikara and Folks [1989].

For example, consider the WSMR base data. We wish to test the hypothesis that $\mu = 1.5$ against the alternate hypothesis that $\mu \neq 1.5$. The test statistic, from Section 9, follows the t–distribution with $n-1$ degrees of freedom. We have 79 data points, so our critical value is $t_{crit} = 1.99045$ at the 0.05 significance level.

We compute the value of the statistic and obtain:

$$t = \left| \frac{\sqrt{n-1}(\overline{X} - \mu_0)}{\mu_0 \sqrt{\overline{X} V}} \right| = 3.032 > 1.99 \tag{4}$$

We reject the hypothesis that $\mu = 1.5$. This accords with the results of our previous section, where 1.5 was not included in our 95% confidence interval for μ, given in Equation 3.

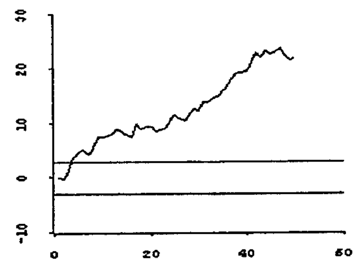


Figure 5: A graphical description of the SPRT.

Sequential Testing

It is possible to test if the means of two combat models are equivalent using sequential methods. In these methods, one does not predetermine the number of simulation runs, but rather samples until one can make a decision. The classic method is the sequential probability ratio test.

Wald conjectured [Wald, 1947] and later proved [Wald and Wolfowitz, 1948] that the sequential probability ratio test (SPRT) is optimal for deciding between two point hypotheses in the sense that the expected number of points sampled before a decision could be reached was minimized with the SPRT. A precise statement of these optimality properties of the SPRT in a decision framework can be found in [Ferguson, 1967].

The SPRT considers

$$\Lambda_n = \frac{f(X_1, X_2, \ldots, X_n|\theta_1)}{f(X_1, X_2, \ldots, X_n|\theta_0)} = \prod_{i=1}^{n} \frac{f(X_i|\theta_1)}{f(X_i|\theta_0)} \tag{5}$$

where $f(x|\theta)$ is the joint or marginal density as appropriate. The SPRT accepts $H_0 : \theta = \theta_0$ if $\Lambda_n \leq A$, accepts $H_a : \theta = \theta_1$ if $\Lambda_n \geq B$ and otherwise continues sampling. This is illustrated in Figure 5, with $A = -3$ and $B = 3$, where the null hypothesis would have been rejected at observation number 4.

In practice, we work with the log-likelihood, or $\ln(\Lambda_n)$, which results in a cumulative sum. We accept, reject, or continue sampling based on the value of this cumulative sum. As we have written it, the log-likelihood ratio will have a negative expected value when the process is in-control. When the process is well modeled by the alternate hypothesis, the log-likelihood ratio will have a positive expected value. As a result, when the process is in-control, the sum tends downward. When the process is out-of-control at the alternative distribution, the sum tends upward. When the sum is above a certain limit, we have evidence in favor of the alternative hypothesis. When the sum is below a certain limit, we decide in favor of the null hypothesis. When the sum is in-between the limits, we continue to sample.

In the present context, we would apply the SPRT as follows. We would first have our estimate of the base case parameters, which would determine θ_0. We would then select the shift in the parameter for which we

desire maximum sensitivity. For example, say our estimate of the mean for the base case was $\mu_0 = 1.69$. Say further that we wished maximum power to detect if the mean had shifted to $\mu_1 = 2.00$. We would construct the SPRT with those two point hypotheses, and sample until we reached a conclusion.

The values of the upper and lower limits for the SPRT are set after considering the desired performance of the test in terms of type I (α) and II (β) errors. Exact methods are available, but the usual approximation is to set $A = \alpha/(1-\beta)$ and $B = (1-\alpha)/\beta$.

By using sequential methods, one is guaranteed to reach a decision, and to do so in the fewest average number of simulations. This avoids the situation where one runs, say, the usual thirty trials, fails to reject the null hypothesis, yet doesn't know if 5 more trials would have resulted in the rejection of the null. This avoids the need to do the power calculations discussed next.

Number of runs

To determine the number of simulation runs necessary to detect a difference of parameter of a given size with a given probability, the usual course is to use the power function for the test. The power functions for the inverse gaussian distribution test statistics are not known, however, because the non-central distributions of the test statistics are intractable. In this section, we sketch an approximate method for determining the number of simulation runs necessary.

We assume that we have historical data on the current model, with summary statistics given by $\overline{X}, V_X$, and n_X. This corresponds to the knowledge we would have about the current model after n_X runs.

First, we need to specify two models and error probabilities: the current model, the smallest model change that we wish to detect, the probability of type 1 error (reject the null when it is true) and the probability of type 2 error (accept the null when it is false).

For example, we could identify our current model as represented by the WSMR base case data. We want the probability that we say incorrectly say that the model has changed, when it remains constant, to be less than 5%. We desire to be 95% sure that we detect a model shift to $\mu = 2.00$, with λ remaining constant. In other words, we want $\alpha = 0.05$, $\beta = 0.05$. How many trials should be run?

Our setup consists of two samples, one known and one to be drawn. Here the known sample is the WSMR data. We want to know how large the sample should be for the one remaining to be drawn.

Under the null hypothesis that the means are equivalent, the distribution of the test statistic T given by Equation **??** is known to have the t distribution. As a result, we can compute our critical value for the test statistic. For the WSMR data, with its large sample size, we can approximate the critical value by 2.00, regardless of the size of the second sample.

Under the alternate hypothesis, $\mu = 2.00$. We can draw samples of size n repeatedly, compute T, and find the approximate probability that $T < t_{crit}$. This gives us an empirical estimate for β, the probability that we don't detect the model shift to $\mu = 2.00$ when it has occurred.

Routines for these simulations are easily implemented. One such LISP implementation is available from

the author.

For example, we return to the WSMR base-case data. How many runs do we need to make to be 95% sure to detect a change this large?

We set $n = 200$. Of a thousand trials, 977 have a value of T greater than 2.00. We set $n = 180$. Then 965 of a thousand trials have a value of $T > 2$. We set $n = 150$, and find 937 of a thousand trials have a value of $T > 2.00$. We could apply a bisection method or a simple interpolation to find that we need to set $n \approx 165$ to achieve our desired design.

We note that these simulations take a few minutes to run on a personal computer, but are much quicker than the corresponding JANUS or CASTFOREM simulations.

We have found the simulation community is generally unaware of the large number of simulation runs required to have high power for hypothesis tests when the underlying distribution is as variable and skew as the distribution of LER.

Examples

We present two short examples to support the ideas in this report. The first data set was provided by Mr. Dave Durda of TRAC-WSMR, and is called the WSMR data throughout this paper. The second was provided by Mr. Tom Herbert of the RAND corporation, and is called the RAND data.

WSMR

TRAC-WSMR is responsible for stochastic combat simulation models. One of their models is CASTFOREM. There has been discussion recently about adjusting the way that CASTFOREM assesses damage to systems represented in the model. One proposal was to model degraded states, where instead of a system having a binary state space ("killed" or "not killed"), the system could take on one of several states representing reduced capability.

Three new types of rules were proposed, along with one base case. We call them the base case, and cases one through three. There was interest in whether or not these different rules affected the performance of CASTFOREM, and if so, by how much.

We were provided with the results of 260 simulations of the different rules in CASTFOREM using a standard scenario. The base case and cases two and three were run 80 times each through the standard scenario. The first base case was only run 20 times. We call the base case data "WSMR", the old rule data "WSMR1", and the two other methods "WSMR2" and "WSMR3".

Boxplots for the LERs from the simulation for each of the models are in Figure 6. We see immediately from the boxplots that the old rules clearly produce different results from the three new rules; the graphical evidence is compelling and sufficient. We move on to the question of whether or not the three new methods produce similar results.

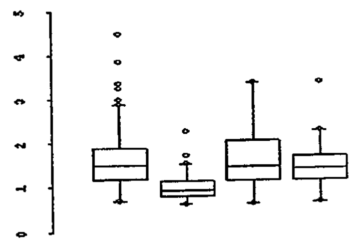


Figure 6: Boxplot of the four WSMR data sets. From left to right, they are the base case, the old binary rules, and two modifications to the base case. Source: White Sands Missile Range, 1996.

The data sets were each found to be well modeled by the IG distribution.

We compute the confidence intervals for the means of WSMR, WSMR2 and WSMR3, and obtain:

$$\mu_{WSMR} \in (1.5635, 1, 858) \tag{6}$$
$$\mu_{WSMR2} \in (1.410, 1.667) \tag{7}$$
$$\mu_{WSMR3} \in (1.3833, 1.6408) \tag{8}$$

It appears from the confidence intervals that the WSMR2 and WSMR3 means are indistinguishable. Can they be distinguished from the base case?

Applying the two sample test developed earlier, we find that WSMR and WSMR2 are not statistically significantly different, as the value of the resulting t statistic is only $t = 1.7468$. The test for equality of means between WSMR and WSMR3, however, has a t statistic value of $t = 2.071$, which is significant at the 0.05 level. We conclude that the WSMR3 set of rules for degraded states has a statistically significantly different impact on LER than the WSMR rules.

We note in conclusion here that if we naively apply the two sample t test which would follow from the inappropriate assumption that WSMR and WSMR3 were normally distributed, or from an asymptotic approximation based on an application of the law of large numbers, we would obtain $t = 1.62$, and **we would not detect the model differences.** Our methods are more powerful than the asymptotic normal approximation.

RAND

We have a second group of data sets, provided by RAND. This data came from trials of the effects of a new weapon system. Three scenarios were run. In the base case, a blue battalion task force attacked a defending red battalion task force. In the second case, the attackers were augmented with a new weapons system. In the third case, the attackers were augmented with two new weapons systems. The simulators sought to demonstrate that the LER was significantly better (from the blue point of view) with the new weapons systems.

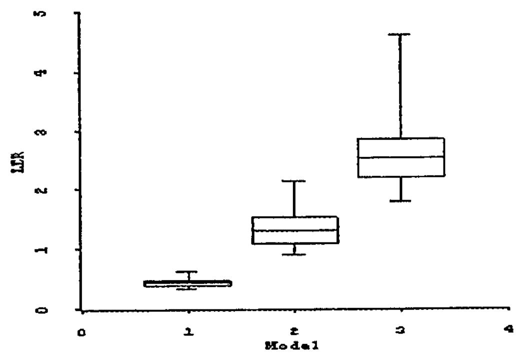


Figure 7: Boxplots for the RAND data. Source: RAND Corporation, 1996.

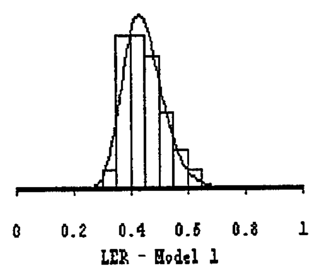


Figure 8: Histogram of the first RAND data set with fitted IG density.

RAND conducted thirty runs of each case.

Boxplots for the three cases are presented in Figure 7. From the boxplots, we see again that no formal statistics are necessary to see that the new weapons systems help the blue force. We can obtain confidence intervals for μ and λ to emphasize the point.

We prefer to dwell on a different point: despite the difference in combat simulations between JANUS and CASTFOREM, both produce distributions of LER which are well modeled by the inverse gaussian distribution. We present some graphical evidence in Figures 8, 9 and 10. Formal testing using Wilks-Shapiro and Kolmogorov-Smirnov tests supports this graphical evidence.

Conclusions and Recommendations

This is a quick summary of the main points of this paper.

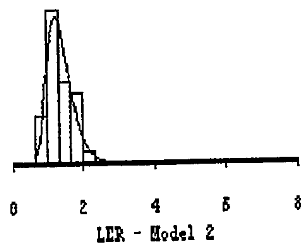


Figure 9: Histogram of the second RAND data set with fitted IG density.

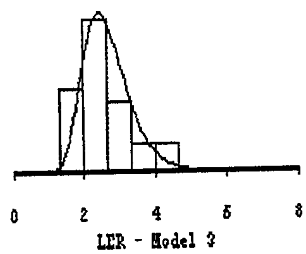


Figure 10: Histogram of the third RAND data set with fitted IG density.

Conclusions

- The inverse gaussian distribution fits LER data well.
- The inverse gaussian distribution provides a complete theory of estimation, hypothesis testing, and design of simulation studies for the use of the analyst. This theory is largely based on standard distributions, such as the t, normal, and χ^2 distributions, which are accessible to analysts.
- Methods based on the inverse gaussian distribution are more powerful for analysis of LER problems than methods based on asymptotic normality.
- Using the methods of this paper, it is possible to easily and accurately approximate the number of simulation runs necessary to detect a change in the mean or shape of the distribution of LER results.

Recommendations

Loss exchange ratios should be modeled as inverse gaussian random variables in studies where high statistical precision is desired.

Further applications of this model should be studied. One promising area is the development of regression models which predict the LER for a given level of JANUS or CASTFOREM parameter associated with some

system capability. This could allow the acquisition community to decide how on a desired level of capability before setting specifications for systems procurement and design. In particular, regression models based on the inverse gaussian distribution with several predictors seem fruitful for future study.

References

[1] Banerjee, Asit K. and G. K. Bhattacharyya (1979) Bayesian Results for the Inverse Gaussian Distribution with an Application. *Technometrics* Vol. 21(2). pp. 247-251.

[2] Bickel, Peter J. and Kjell A. Doksum. (1977) *Mathematical Statistics.* Englewood Cliffs, NJ: Prentice Hall.

[3] Chhikara, R. S. (1975) Optimum tests for the comparison of two inverse Gaussian distribution means. *Australian Journal of Statistics.* Vol. 17. pp. 77-83.

[4] Chhikara, R. S. and J. L. Folks. (1976) Optimum test procedures for the mean of first passage time in Brownian motion with positive drift (inverse Gaussian distribution). *Technometrics.* Vol. 19. pp. 189-193.

[5] Chhikara, R. S. and J. L. Folks. (1977) The Inverse Gaussian Distribution as a Lifetime Model. *Technometrics* Vol. 19. No. 4. pp. 461-468.

[6] Chhikara, R. S. and J. L. Folks. (1989) *The Inverse Gaussian Distribution.* New York: Marcel Dekker.

[7] Chhikara, Raj S. and Irwin Guttman. (1982) Prediction Limits for the Inverse Gaussian Distribution. *Technometrics* Vol. 24. No. 4. pp. 319-324.

[8] Desmond, A. F. and G. R. Chapman. (1993) Modeling Task Completion Data with Inverse Gaussian Mixtures. *Applied Statistics.* Vol. 42. No. 4. pp. 603-613.

[9] Dupuy, Trevor N. (1987) Can We Rely Upon Computer Combat Simulations? *Armed Forces Journal International.* August. pp. 58-63

[10] Edgeman, Rick L., Robert C. Scott, and Robert J. Pavur. (1988) A modified Kolmogorov-Smirnov Test for the Inverse Gaussian Density with Unknown Parameters. *Communications in Statistics - Simulations.* Vol. 17. No. 4. pp. 1203-1212.

[11] Ferguson, Thomas S. (1967) *Mathematical Statistics: A Decision Theoretic Approach.* New York, Academic Press.

[12] Folks, J. L and R. S. Chhikara (1978) The Inverse Gaussian Distribution and its Statistical Application – A Review. *Journal of the Royal Statistical Society B.* Vol. 40. No. 3. pp. 263-289.

[13] Fries, Arthur. (1986) Optimal Design for an Inverse Gaussian Regression Model. *Statistics and Probability Letters.* Vol. 4. pp. 291-294.

[14] Geisser, Seymour (1993) *Predictive Inference: An Introduction* New York: Chapman & Hall.

[15] Helmbold, Robert L. (1990) *Rates of Advance in Historical Land Combat Operations.* Bethesda, Maryland: US Army Concepts Analysis Agency.

[16] IMSL, Inc. (1989) *MATH/LIBRARY: FORTRAN Subroutines for Mathematical Applications.* Edition 1.1. Houston: IMSL, Inc.

[17] Hughes, Wayne P., editor. (1984) *Military Modeling.* Alexandria, VA: Military Operations Research Society.

[18] Lanchester, F. W. (1956) Mathematics in Warfare. In *The World of Mathematics.* Vol. 4. Edited J. R. Newman. New York: Simon and Schuster. pp. 2138 -2157.

[19] Olwell, David H. (1996) *Topics in Statistical Process Control* Ann Arbor: University Microfilms.

[20] Savage, I. R. (1962) Surveillance Problems. *Naval Research Logistics Quarterly.* Vol. 9.

[21] Schroedinger, E. (1915) Zur Theorie der fall- und steigversuche an teilchen mit Brownscher bewegung. *Phys. Ze.* Vol. 16. pp. 289–295.

[22] Taylor, H. M. (1965) Statistical Control of a Gaussian Process. *Technometrics.* Vol. 9.

[23] Taylor, James G. (1981) *Force-on-Force Attrition Modeling.* Arlington, VA: Operations Research Society of America.

[24] Taylor, James G. (1983) *Lanchester Models of Warfare.* Volumes I and II. Arlington, VA: Operations Research Society of America.

[25] Tierney, Luke. (1990) *LISP-STAT: An Object Oriented Environment for Statistical Computing and Dynamic Graphics.* New York: Wiley.

[26] Tweedie, M. C. K. (1957a) Statistical properties of inverse Gaussian distributions I. *Annals of Mathematical Statistics.* Vol. 28. pp. 362-377.

[27] Tweedie, M. C. K. (1957b) Statistical properties of inverse Gaussian distributions II. *Annals of Mathematical Statistics.* Vol. 28. pp. 696-705.

[28] Varian, Hal R. (1975) A Bayesian Approach to Real Estate Assessment. In *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage* Eds. Stephen Feinberg and Arnold Zellner. Amsterdam: North-Holland. pp. 195-208.

[29] Ventisel, Ve. S. (1964) *Introduction to Operations Research* Moscow: Soviet Radio Publishing House.

[30] Wald, Abraham (1945) Sequential Tests of Statistical Hypotheses. *Annals of Mathematical Statistics.* Vol. 16.

[31] Wald, Abraham. (1947) *Sequential Analysis* New York: Wiley.

[32] Wald, Abraham, and Jacob Wolfowitz. (1948) Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics.* Vol. 19. pp. 326-339.

[33] Zellner, Arnold. (1986) Bayesian Estimation and Prediction Using Asymmetric Loss Functions. *Journal of the American Statistical Association.* Vol. 81. No. 394. pp. 446-451.

INTENTIONALLY LEFT BLANK.

CASTFOREM VERIFICATION AND VALIDATION PROCESS

Douglas C. Mackey
TRADOC Analysis Center-White Sands Missile Range
White Sands Missile Range, New Mexico 88002-5502

ABSTRACT

This paper describes the past, present, and future verification and validation (V&V) efforts for the Combined Arms and Support Task Force Evaluation Model (CASTFOREM). CASTFOREM is the Army's brigade level high resolution land combat simulation model. It has been used in numerous studies and cost and operational analyses and, as such, has undergone an elaborate verification and validation of its data and algorithms.

The generalized verification and validation processes will be discussed, specific examples will be provided, and then a history of all efforts will be listed.

The ability to simulate reality is a challenge that may never be met but will always be a goal. CASTFOREM strives to meet the challenge by using a continuous V&V process. The summation of many V&V efforts, over many years of use, have earned CASTFOREM a high degree of credibility in the army modeling community.

INTRODUCTION

CASTFOREM models all types of direct fire, crew-served ground weapons systems; helicopters; dismounted infantry (fire teams); artillery (ICM, guided munitions, smart munitions, smoke); engineering operations (minefields, barriers, and breaching); combat service support functions (rearm, refuel); communications (including networks); maneuver with capability of dynamic route selection; detailed search and acquisition (multiple sensors using Night Vision and Electronic Sensors Directorate (NVESD) modeling); and realistic battlefield (smoke, dust, weather, Army Research Laboratory-Battlefield Environment Directorate's (ARL-BED) COMBIC model; digitized terrain). CASTFOREM is highly flexible both as to what it can model and as to the degree of resolution to which an object or process is modeled.

Each organizational entity (commanders and units of resolution, e.g., tanks, infantry fighting vehicles (IFV), and trucks) possesses a singular intelligence system which is updated by the acquisition of information via a communication net or directly (detecting a target, encountering an obstacle, receiving fire, etc.). Delays and failures in the exchange of information over a communication net will cause each entity's intelligence system to perceive battlefield knowledge rather than perfect knowledge. The latter, however, can be represented by simulating perfect and instantaneous exchange of information among organizational entities.

In general, all combat support and combat service support units and functions which interact with and/or directly affect the combat activities of maneuver units are represented in the model. The degree of resolution to which all units and their functions are modeled is greatest for maneuver units, less for combat support units, and least for combat service support units. However, the CASTFOREM structure facilitates increasing the degree of resolution with which specific vehicles, weapons, and functions are represented to satisfy user study objectives.

The CASTFOREM scenario preparation process closely parallels the military planning process for a tactical operation in terms of methodology. This is accomplished through the construction of knowledge bases (via decision tables) for both Red and Blue. Each knowledge base is designed for a specific type tactical operation (e.g., active defense, deliberate attack, hasty river crossing); contains doctrinal responses to a broad spectrum of tactical situations; requires user threshold inputs to trigger each doctrinal response; and permits dynamic maneuver by opposing forces.

CASTFOREM is comprised of the following process modules:

- Command and Control (C2)
- Communications (COMMO)

Approved for public release; distribution is unlimited.

- Combat Service Support (CSS)
- Engineer (ENGR)
- Surveillance (SEARCH)
- Engage
- Maneuver
- System/Environment

The model contains the C2 (inference engine) logic, which accesses the knowledge base to make tactical decisions which generate orders, reports, and requests for support. In turn, these decision table outputs control the actions of units of resolution. This logic, combined with explicit representation of a C2 structure and communication nets, represents the C2 process employed by combat units.

The resolution of CASTFOREM is at the individual vehicle (e.g., a tank, an APC, or a truck) or individual soldier and there are no artificial limits on the sizes of the forces played. Usual battle times run from 30 minutes to 3 hours.

Figure 1 portrays the fundamental cycle of integration over time for each CASTFOREM unit.

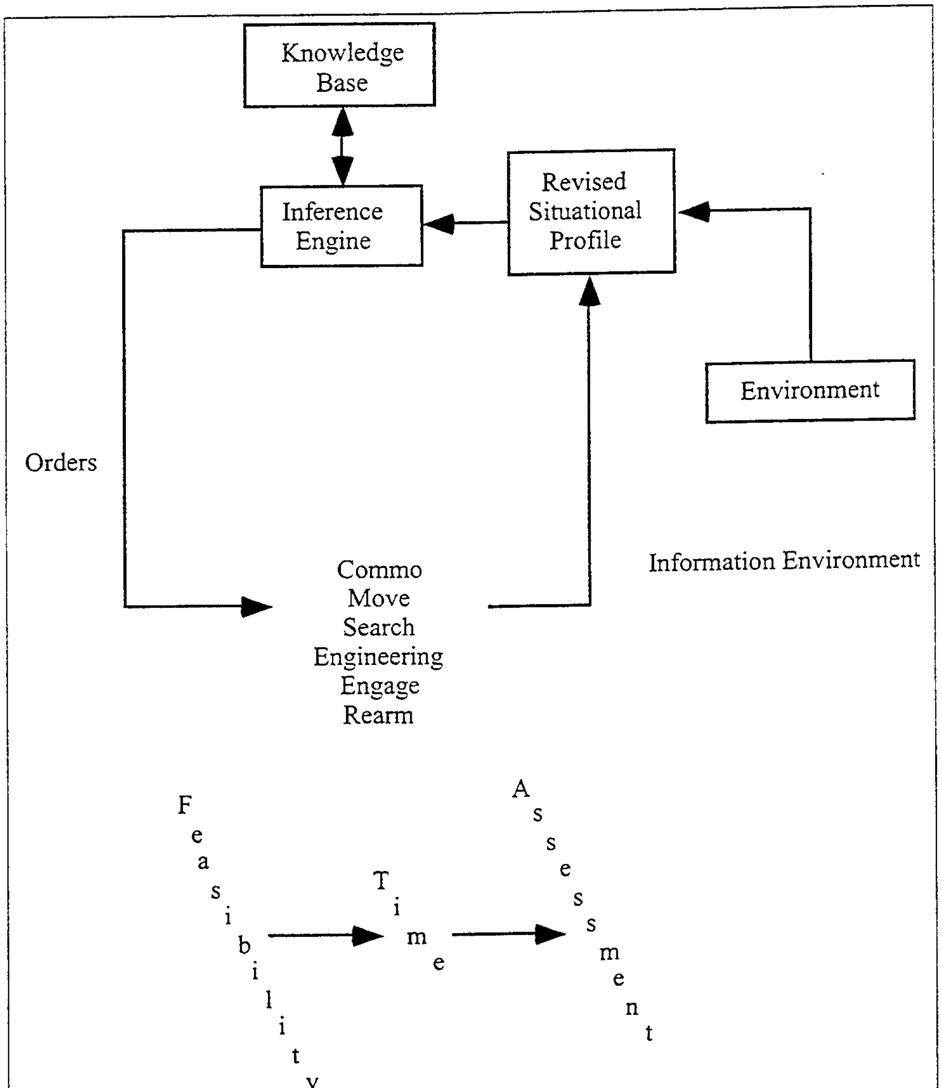


Figure 1. Fundamental Cycle of Integration

Initially, each unit will receive their first combat orders. They may direct the unit to move, search, communicate, etc. The unit determines if it is feasible to execute the order and, if so, schedules an event for its completion. After time has been charged, an assessment is computed as to the event completion (e.g., reached a destination) and a decision table may be executed to determine the next appropriate order.

GENERALIZED VERIFICATION AND VALIDATION PROCESS

The definition of V&V from AR 5-11, the technique of V&V process, and configuration control for the CASTFOREM are described.

In accordance with AR 5-11, we have the following definitions:

Verification, in the context of this regulation and Army Model Improvement Program (AMIP) models, is defined as a technical review of a model's algorithms to ensure their suitability for the model's intended purpose. Such a review must be designed to determine if algorithms are technically sound, consistent with current approved analytical techniques, and appropriate to the model design.

Validation, in the context of AR 5-11 and AMIP models, refers to an iterative process designed to determine whether the model/simulation reflects results expected in the real world. It must be recognized that, due to the complex nature of the real world, no validation effort can be expected to be totally accurate. Nonetheless, by approaching validation through a logical sequence of iterative steps (outlined in paragraph below), an evaluation of a model's approximation of reality can be obtained.

Verification and validation are indeed a continuous process over time. For CASTFOREM, every time a new algorithm is introduced it undergoes a V&V process to insure it is "reasonable."

The verification process is fairly straight forward as outlined below. It is validation that is difficult.

Complicating the validation effort are the following:

- quantifying human factors for input to the model
- modeling futuristic weaponry
- benchmark field tests may not represent actual battlefield conditions
- historical data from actual battles may not be representative of future battles

Absolute validity will never be achieved but always remains as a goal for CASTFOREM.

To help facilitate algorithmic verification, validation, and consistency, reference (3) was published by AMSAA. This compendium of high resolution attrition algorithms describes CASTFOREM's algorithms in detail.

Here are some thoughts from references 1 and 2 all of which apply to CASTFOREM:

"Without *validation* a model is of very little use. The concepts of *inductive* and *deductive* reasoning are introduced, and it is shown that it is impossible to validate models in the strictest sense of the word. Modeling is not a precise science; hence the criteria used for testing the robustness of scientific theories should not be strictly applied to models at the present time. Here, 'validation' means substantiating that the model within its domain of applicability is sufficiently accurate for the intended applications. The emphasis is on establishing the degree of confidence in the model rather than testing for its absolute validity, and this is achieved by collecting evidence to support the validity of concepts, methodology, data, experimental results, and inference. Model sponsors, model builders, and model users must be prepared to accept compromise solutions."[1]

"The ease or difficulty of the validation process depends on the complexity of the system being modeled and on whether a version of the system currently exists. For example, a model of a neighborhood bank would be relatively easy to validate since it could be closely observed. On the other hand, a model of the effectiveness of a naval weapons system in the year 2025 would be virtually impossible to validate completely,. since the location of the battle and the nature of the enemy weapons would be unknown."[2]

"A simulation model of a complex system can only be an *approximation* to the actual system, regardless of how much effort is put into developing the model. There is no such thing as an absolutely valid model."[2]

"A simulation model should be validated relative to those measures of performance that will actually be used for decision making."[2]

"... model development and validation should be done hand-in-hand throughout the entire simulation study."[2]

VERIFICATION PROCESS

Data verification techniques are:

- Identification of data

- Traceability of data to approved sources (e.g., Ballistics Research Laboratory (BRL), Army Materiel Systems Analysis Activity (AMSAA), etc.)

- Analysis of the use of the data

Algorithm verification techniques consist of:

- Running parametric sensitivities on the algorithm in a stand-alone environment and as an integral part of CASTFOREM and then analyzing the outputs vis-à-vis the inputs. Output is analyzed from a numerical, statistical, and behavioral perspective to determine if the first, and higher order effects of the algorithm have surfaced as intended by the modeler.

- Structured walk-through of the stand-alone algorithm and its model interfaces. This technique enables all personnel involved to come to the same plane of understanding regarding expected model outputs. It provides the opportunity for the designer, coder, and reviewer to make a detailed review of the coded algorithms' structures to ensure that the algorithm functions as intended by the modeler and that the necessary dynamic data interactions take place properly.

- Day-to-day checkout of code by the programmers.

- Day-to-day checkout of scenarios by the analysts.

- Briefing new algorithms at annual CASTFOREM users' group meeting.

VALIDATION PROCESS

Data validation is accomplished by ensuring that the data be representative of some empirical standard to attain consistency and reasonableness. AMSAA is key to this.

Algorithm validation is accomplished by, once again, running parametric sensitivities on the algorithm in a stand-alone environment and as an integral part of CASTFOREM. Then, one or several of the following techniques are applied to ensure "reasonableness" and consistency of the model output:

- Field test comparisons

- Comparison of results to historic data or other model results (benchmarking)

- Independent review, either by designated committees or by functional area experts, to determine if model results are "reasonable"

- Study advisory groups (SAGs)

- Peer review groups within the study process

CONFIGURATION CONTROL

Frequently, new algorithms or updates to old algorithms are presented for potential integration into CASTFOREM. This section describes the process by which changes are brought into CASTFOREM.

Figure 2 portrays CASTFOREM configuration control. In general, modifications are desired by a user, either local or remote. Those modifications are coded and checked out in a test environment. Once they are ready for integration into CASTFOREM, TRADOC Analysis Center-White Sands Missile Range (TRAC-WSMR) begins a V&V effort.

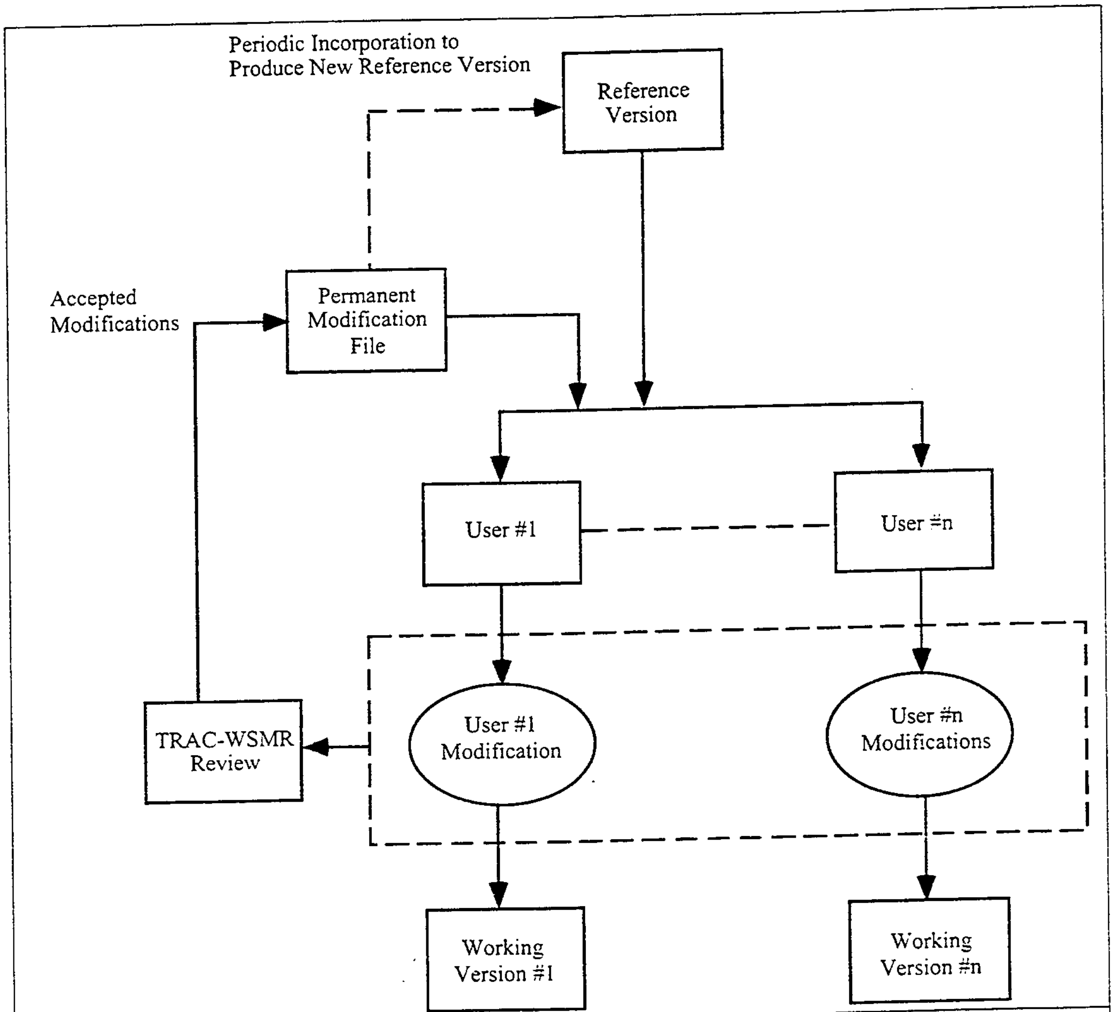


Figure 2. CASTFOREM Configuration Control

Figure 3 portrays TRAC-WSMR's V&V review process.

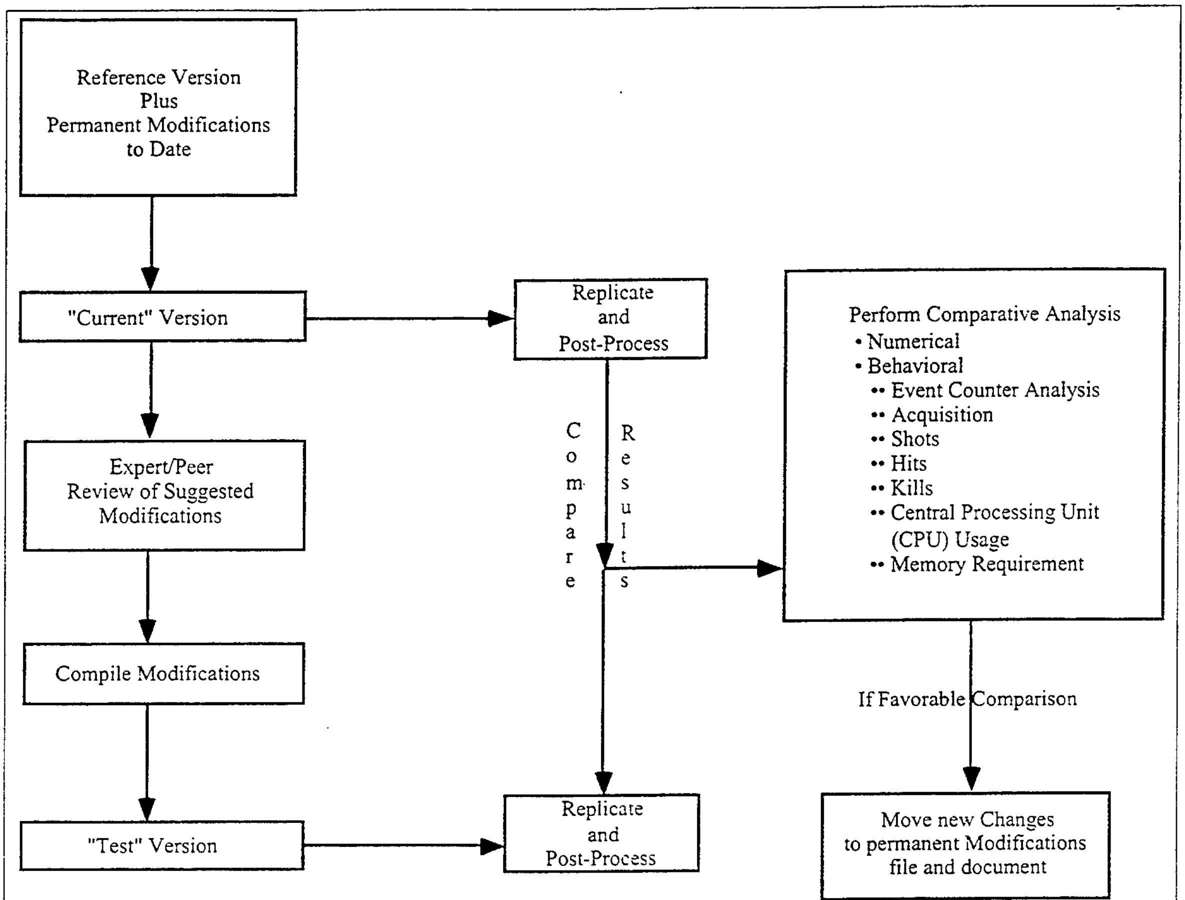


Figure 3. TRAC-WSMR's V&V Review Process

First the "current" version of the model is replicated and post-processed with a benchmark scenario.

Second, the new code is compiled into the model providing a "test" version. The test version is then replicated and post-processed.

Third, the results of each set of runs is compared. If the comparison is favorable, the new code is moved to the permanent modification file and documented.

The algorithm and its supporting data undergo all applicable/possible V&V efforts described above. Once it is agreed that the new algorithm/data produce the correct effect, the code is integrated into CASTFOREM.

Algorithm integration into CASTFOREM consists of:

- Documenting the routines added/modified
- Saving the old code as backup
- Moving the new code to the reference version

VERIFICATION AND VALIDATION EXAMPLES

EXAMPLE 1. CASTFOREM COMPARISON TO CARMONETTE MODEL

CARMONETTE was the previous high resolution model used widely by the Army community. It had gained a high degree of credibility over the years and was considered the benchmark simulation.

Benchmarking CASTFOREM against CARMONETTE was an expedient way of "inheriting" some of CARMONETTE's credibility.

The Armor Investment Strategy (AIS) Study was chosen as the first benchmark. This study had already been run in CARMONETTE with three main alternatives: Blue tank lethality, Blue tank accuracy, and ITOW. It was then rerun using CASTFOREM. The findings of the study were the same.

More importantly, not only did CASTFOREM provide end game statistics comparable to CARMONETTE, but Blue and Red losses over time were also comparable. See figure 4. This showed that the battle evolved over time in the same way in both models. This was a major milestone for CASTFOREM.

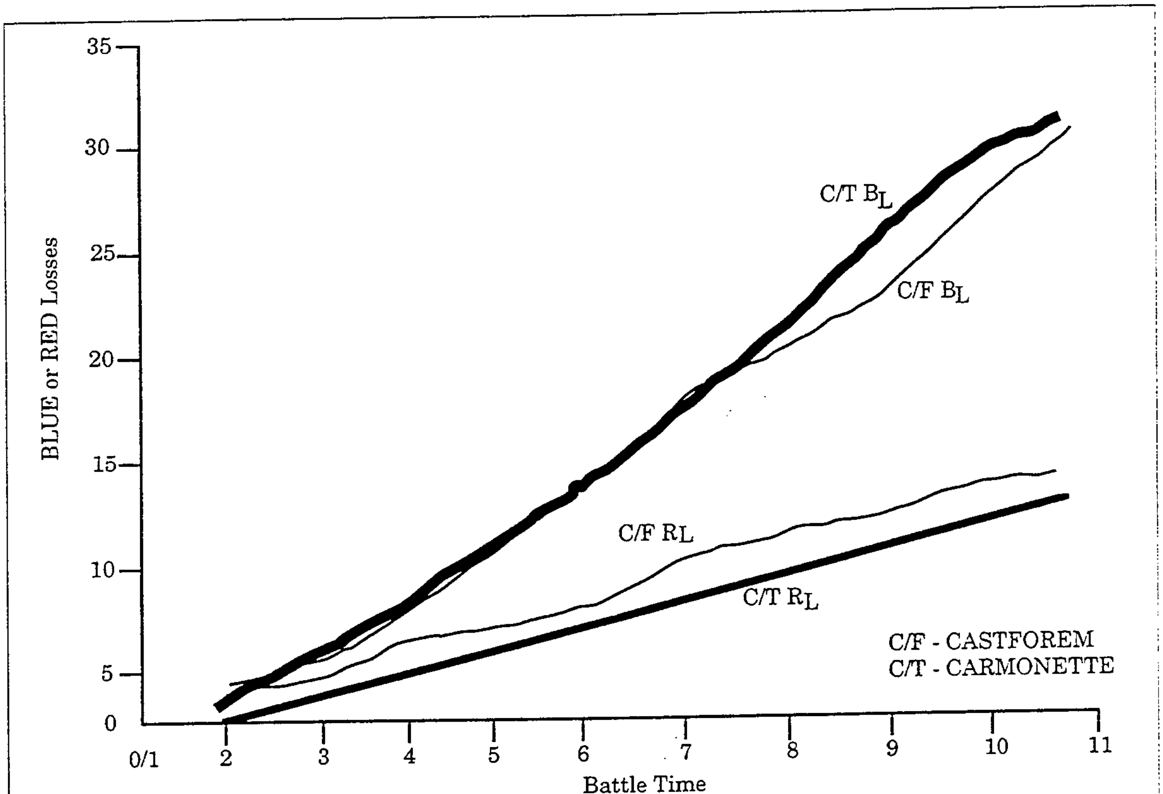


Figure 4. Comparison to CARMONETTE Model - 1985

EXAMPLE 2. FAADS MODEL-TEST-MODEL

This was a large effort to compare the pedestal mounted Stinger (PMS), MANPADS Stinger teams, and LOS-F-H to the field. Day, night, MOPP, NOMOPP, CUED, and AUTONOMOUS cases were all run. The ranges of detection, engagements, and intercepts were compared to the field.

In general, intercept and engagement ranges compared favorably with the field. Detection ranges did not. The collection of detection ranges from the field was difficult. Also, the NVL detection model produced detections at much shorter ranges than in the field. Some sample results for PMS and MANPADS are provided in figure 5.

Pedestal Mounted Stinger								
			Ranges					
			Detections		Engagements		Intercepts	
			Fixed-Wing	Rotary-Wing	Fixed-Wing	Rotary-Wing	Fixed-Wing	Rotary-Wing
Day	MSCS	MOPP0	S	S	N	S	S	S
Day	Auto	MOPP4	S	S	N	N	N	N
Day	MSCS	MOPP4	S		S		N	
Day	Auto	MOPP0	S	N	N	S	N	N
Day	MSCS	MOPP0	N		S		S	
Day	Auto	MOPP0	N		N		N	
Night	MSCS	MOPP0	S		S		S	
Night	Auto	MOPP4	N		N		N	
Night	RDDS	MOPP4	S		N		S	
Night	Auto	MOPP0	S		N		N	
Day	RDDS	MOPP0	S	N	N	S	N	S
Day	RDDS	MOPP0	N	S	N	S	N	N
Day	RDDS	MOPP4	N		N		N	
S - Significant Difference From Field					N - No Significant Difference			

MANPADS								
			Ranges					
			Detections		Engagements		Intercepts	
			Fixed-Wing	Rotary-Wing	Fixed-Wing	Rotary-Wing	Fixed-Wing	Rotary-Wing
Day	MSCS	MOPP0	S	S	N	N	N	N
Day	Auto	MOPP4	S	S	N	N	S	N
Day	MSCS	MOPP4	S		N		N	
Day	Auto	MOPP0	S		N		N	
S - Significant Difference From Field					N - No Significant Difference			

Figure 5. Model-Test-Model Effort

This effort won the Wilbur B. Payne award.

EXAMPLE 3. SIMNET-D M1A2 SYNTHETIC ENVIRONMENT EXPERIMENT

This was an experiment using SIMNET-D and man-in-the-loop M1A2 simulators.

Several modeling insights were gained. CASTFOREM implemented implicit and explicit target cueing, a degradation in the usage of CITV (from 100 percent previously), reorienting the hull toward the enemy to increase survivability, and new tank gunner disengage logic. This logic would disengage a tank gunner only when the target stops moving and firing.

EXAMPLE 4. LONGBOW IOTE LINKAGE TO COEA (1995)

This is one of the most recent efforts. It compared CASTFOREM results to a similar scenario flown in the IOTE.

Similarities were shown in the percentage increase of loss exchange ratios, number of kills per system, survivability between basecase and longbow, and helicopter tactical sequences.

Timelines did not compare favorably and this is still an open question.

VERIFICATION AND VALIDATION EFFORTS

VERIFICATION

Reference 2 provides several techniques for verification which have been used extensively in CASTFOREM. (There are many good references on V&V. I choose 1 and 2 as representative.)

- Technique 1: "In developing a simulation model write and debug the computer program in modules or subprograms."

This was a coding convention imposed on the original coding team of CASTFOREM and remains in force today. There are nine major modules of code: surveillance, maneuver, combat service support, engineer, engage, communications, command, control, and the system. Each of these, in turn, have numerous submodules.

- Technique 2: "It is advisable when developing large simulation models to have more than one person read the computer program."

During past and current development of CASTFOREM, any coding done by a team member was always reviewed by the chief programmer, at a minimum. Over the past several years, due to employment turn over, various modules of code have been passed to new team members who, in turn, begin by flow charting the module. This has provided an excellent "second look".

- Technique 3: "The model should be run under simplifying assumptions for which its true characteristics are known."

Whenever a new algorithm is being tested in CASTFOREM, it is analyzed using a one-on-one or few-on-few scenario that is tailored to the new coding. This allows the programmer a chance to compare the result to a hand calculated result. Then it is tested out further in a many-on-many scenario. Finally, it is tested on several large "high resolution" scenarios.

- Technique 4: "With some types of simulation models, it may be helpful to observe an animation of the simulation output."

This is a very important part of verification for CASTFOREM. CASTFOREM has an elaborate playback capability. It allows a user/coder to visually look at a playback of a battle to determine its overall integrity.

VALIDATION

Reference 2 provides a three step approach for developing a valid and credible simulation model. The three steps include:

1) Developing a high face validity via expert review, peal backs, briefings, and comparisons to already accepted models.

2) Empirical testing of model assumption using sensitivity analysis.

3) Comparing model output to field test data.

Table I is a chronological listing of all major algorithmic V&V efforts for CASTFOREM.

Table II is a listing of all major studies CASTFOREM has successfully completed. some of the major combat simulations that CASTFOREM has been compared to in some detail.

Table III outlines table IV lists all field tests and battles that CASTFOREM has been compared against.

As one can see, CASTFOREM has, over the years, undergone an enormous amount of V&V which continues on a daily basis. As a consequence, CASTFOREM has earned a high degree of credibility in the Army modeling community.

Table 1. CASTFOREM Algorithm V&V Efforts

		V&V Efforts				
DATE	ALGORITHM	FACE VALIDITY EXPERT/BRIEF	SENSITIVITY STUDY	USED IN STUDY/COEA	COMPARISON TO FIELD TEST	OTHER COMMENTS
87	CITV	Ft Knox	Yes	MIA Others	M1A2 IOTE	Compared to CARMONETTE
87	Within Burst Dispersion	AMSAA	–	M1A2 Others	–	–
87	Kinematic Boundaries	AMSAA/ Accreditation Brief to SARD	–	FAADS Others	–	Compared to CARMONETTE
87	IRST	AMSAA	–	FAADS	–	Compared to CARMONETTE
88	'Awareness' Range for Dismounted in Foxhole	Ft Benning/ IAAWS Brief	Yes	IAAWS Others	–	Ft Benning Confirmed Range
88	ATR	Ft Rucker AMSAA	–	LHX	–	Documented
88	Bombing	AMSAA/ JMEMS	–	FAADS Others	–	–
88	Fendrikov Sheafing	Ft Sill AMSAA	Yes	All	–	Compared to CARMONETTE (Documented)
88	Artillery Assessment vs Dismounted	AMSAA/ Ft Benning IAAWS Brief	–	IAAWS Others	–	–
89	Use of WES Mobility Data	WES/ Dr. Marcuson III	Yes	All	–	Documented
89	Artillery Module	Peelback	–	Most	Study of Artillery Effects	Compared to CARMONETTE, VIC, TAFSM, Janus (Documented)
89	Foxhole Geometry	Ft Benning	Yes	SAMP	–	Described by Infantry Soldiers
89	Dynamic Dimensions	AMSAA	–	ASM Others	–	Allows LOSAT to change from STOWED to DEPLOYED
89	SADARM	AMSAA	–	ASM Others	Captive Flight Test Data	Derived from 'Games' Algorithm (Documented)
89	Vehicle Dust	ARL-BED	Yes	ASM Others	COMBIC Benchmarks	Uses COMBIC Model (Documented)
89	RWR, LWR, Missile Detect Devices	AMSAA	–	ASM Others	–	Documented

Table 1. CASTFOREM Algorithm V&V Efforts (Continued)

		V&V Efforts				
DATE	ALGORITHM	FACE VALIDITY EXPERT/BRIEF	SENSITIVITY STUDY	USED IN STUDY/COEA	COMPARISON TO FIELD TEST	OTHER COMMENTS
89	Laser Rangefinder	AMSAA NVL	YES	M1A2 Others	–	Used to Discriminate 1.06 vs 10.6 (Documented)
89	Counter Battery Radar	Ft Sill	–	All	–	Derived from TAFSM Algorithm
89	WAM	ARDEC AMSAA	–	WAM COEA Others	WAM IOTE	Documented
89	Volcano Conventional Mine Emplacement	AMSAA	Yes	WAM COEA Others	Derived from field test data	–
89	External Event Movement from Field Tests	OPTEC TRAC-WSMR	Yes	FAADS M-T-M Other M-T-M	Mimicks movement in field test precisely	-
89	Air Defense Radar Jamming	AMSAA Misc Ft Rucker	Yes	LHX Longbow Other	–	Table look up based on J/S (Documented)
89	Longbow FCR and Missile	AMSAA MICOM Ft Rucker SARD	Yes	Longbow Others	–	–
89	'Tracking' Boundaries for Air Defense Radars	AMSAA Misc SARD	Yes	LHX Longbow Others	–	–
90	Grenade Smoke and MMW Smoke	Ft Knox AMSAA	–	ASM Others	–	Uses COMBIC
91	Suppression	Ft Sill Briefed at Ft Sill	Yes	Legal Mix VII Others	–	Accepted by Ft Sill
91	Elliptical Carleton Damage Function	AMSAA	Yes	Legal Mix VII Others	Study of Artillery Effects	Documented
91	MLRS TGW	AMSAA	Yes	Legal Mix VII	–	Derived from 'Games' Algorithm (Documented)
91	Towed Howitzer	AMSAA Ft Sill	–	Legal Mix Vii	–	Ft Sill provided degradation due to partial attrition of crew
92	Herring Bone Formation, NCTR, Jamming, Fusion, IFF, Visual ID, C^3I Cues	MICOM AMSAA Ft Bliss Accreditation Brief to SARD	Yes	FAADS COEA Others	Some data from field tests	-

Table 1. CASTFOREM Algorithm V&V Efforts (Continued)

		V&V Efforts				
DATE	ALGORITHM	FACE VALIDITY EXPERT/BRIEF	SENSITIVITY STUDY	USED IN STUDY/COEA	COMPARISON TO FIELD TEST	OTHER COMMENTS
92	MOPP IV Degradation	Chemical School	Yes	NBCRS	CANE II B	Documented
92	Volcano Emplacement of WAM	Engineer School	Yes	WAM COEA Others	–	–
92	Active Protection System	AMSAA	Yes	LOSAT Others	–	Documented
92	Copperhead I Upgrade	Ft Sill AMSAA	Yes	Light weight laser designator COEA	–	–
93	Fratricide	AMSAA	Yes	BCIS COEA Others	–	Increased run time
93	Area Fire	Ft Benning	Yes	Land Warrior	–	Ft Benning described algorithm
93	COMBIC '92 Upgrade	ARL-BED	Yes	All	–	Documented
93	Longbow Upgrade	AMSAA	Yes	Longbow Counter-measure Study Others	–	AMSAA actually did coding and study
93	2 Dimensional MRC/ MRT Model	NVL AMSAA	Yes	Second Generation FLIR Others	–	Critical model upgrade (Documented)
93	Counterbattery vs NLOS	PM-NLOS	Yes	NLOS Vulnerability Study	–	Reimbursable work
93	Dynamic Rearm Logic for Crusader and FARV	Ft Sill	Yes	AFAS COEA Others	–	–
94	Correlate Aiming Errors Within an Artillery Mission	AMSAA	Yes	All	–	Upgrade
94	New Table Look Up Lockon for Longbow Missile	AMSAA	Yes	All	–	AMSAA did coding and checkout (Documented)
94	-CLAYMORES -Body Armor -Grenades vs Dismounted -Dedicated Wpn Algorithm	AMSAA Ft Benning	Yes	Land Warrior	–	Documented

Table 1. CASTFOREM Algorithm V&V Efforts (Continued)

		V&V Efforts				
DATE	ALGORITHM	FACE VALIDITY EXPERT/BRIEF	SENSITIVITY STUDY	USED IN STUDY/COEA	COMPARISON TO FIELD TEST	OTHER COMMENTS
94	C/D Longbow Helicopter Tactics	Ft Rucker AMSAA	Yes	All	–	Warrant Officer described tactics
94	Merlin Mortar	ARDEC	Yes	Mortar Study	–	Reimbursable
94	Explicit False Targets for SADARM	AMSAA	Yes	SADARM COEA Others	Based on field test data	Used poisson distribution (Documented)
94	Vehicle Smoke	ARL-BED	Yes	All	COMBIC Benchmarks	Played through COMBIC (Documented)
94	Missile Countermeasure Device	ARL-SLAD TACOM	Yes	Some	–	Documented
94	COMBIC '93 Upgrade	ARL-BED	Yes	All	COMBIC Benchmarks	Documented
94	Dynamic Sky-over-ground ratio	ARL-BED	Yes	Non	–	Uses Delta Eddington (Documented)
94	NVL Detect Time Calculation Upgrade	NVL	Yes	All	–	Documented
94	New Longbow False Target Algorithm	AMSAA MICOM	Yes	All	Based on test data	–
94	-New TOW and Staff Flyout -New Tanker Disengage Logic -Dynamic Dispersion -Max/Min Caps on Aim Times	AMSAA	Yes	A^2ATD Others	–	Required more detail to compare to ModSAF
94	Laser Range Finder Upgrade	MIT	Yes	BCIS Other	–	Compute % beam on target more accurately
94	Dynamic MOPP IV	Chemical School	Yes	NBCRS	–	–
94	Laser False Target Generator	TACOM	Yes	Guardian Task Force Others	–	Documented
94	Engagement Situational Awareness	PM-CID AMSAA MIT	Yes	Sensitivity only	TIREM Propagation Model Benchmarks	Documented

Table 1. CASTFOREM Algorithm V&V Efforts (Continued)

		V&V Efforts				
DATE	ALGORITHM	FACE VALIDITY EXPERT/BRIEF	SENSITIVITY STUDY	USED IN STUDY/COEA	COMPARISON TO FIELD TEST	OTHER COMMENTS
95	Determining a Hit Using BIAS and Dispersion Data	AMSAA	Yes	All	–	AMSAA verified CASTFOREM output by comparing to their alone model
95	Anti-Helo Mines		Yes			Reimbursable
95	BCIS DDL and TIREM	PM-CID AMSAA MIT	Yes	BCIS DDL Study	–	Results compared favorably to PM's
95	Interpolation of Thermal Contrast When Vehicle Between Hull Defilade and Fully Exposed	AMSAA	Yes	All	–	–
95	Missile Breaklock Algorithms	PEO-Tactical Missile AMSAA MICOM	YES	AMS-H Quick Action Others	Some test data	Peelback of code was done
95	C2 Heuristic for Fire Discipline Within Platoon in Defense	Ft Benning	Yes	AMS-H Quick Reaction Others	–	Described by Ft Benning
95	Correct Adaptation of Javelin Trajectory to Terrain	TI	Yes	All	–	TI found problem by running and analyzing output CRADA
95	Longbow Update	AMSAA	Yes	Longbow COEA Others	Yuma APG IOTE Data	Major upgrade using field test data for COEA
95	'X'-Pattern Emplacement of WAM for area Denial	Engineer School	Yes	A^2R^2 Others	–	Documented
95	Artificial Illumination	ACQSIM AMSAA NVL	Yes	Longbow Others	–	Documented
95	Air Defense Surveillance Radar Detect Model	MICOM AMSAA	Yes	Longbow Others	Some Test Data	Documented Major upgrade using signal to noise + clutter dynamic RCS
95	Cumulative Damage Heuristic (2 Kills = K-Kill)	AMSAA	Yes	A2R2 Others	–	Still under review

Table 1. CASTFOREM Algorithm V&V Efforts (Continued)

		V&V Efforts				
DATE	ALGORITHM	FACE VALIDITY EXPERT/BRIEF	SENSITIVITY STUDY	USED IN STUDY/COEA	COMPARISON TO FIELD TEST	OTHER COMMENTS
96	Frequency Hopping Random & Priority Net Access	AMSAA CECOM Peelback	Yes	Task Force XXI Others	Task Force XXI test at Ft Hood	Benchmark in progress
96	HPM	ARL	Yes	A2R2	–	Documented
96	Intelligent Minefield	AMSAA	Yes	A2R2 Others	–	–
96	'Fire Back' Logic for Dismounted Firing at Dismounted	Ft Benning	Yes	21th Century Warrior	–	Ft Benning described tactic
96	Upgrade to MLRS TGW	Ft Sill AMSAA	Yes	A2R2 Others	–	Changed counter logic, shifted footprint, used oblique attack angle for submunition
96	JAVELIN Minimum Trackable Temperature	TI AMSAA	Yes	Future	–	TI provided and AMSAA accredited
96	MIS-ID	AMSAA		Future		
96	Laser Backscatter	AMSAA		Future		
96	Explicit False Targets for DVO/FLIR			Future		
96	Tactical Internet	CECOM AMSAA		Future		
96	MOUT Operations	Ft Benning		Future		
96	Information Based Decision Model	MIT AMSAA PM-CID		Future		

Table II. Past CASTFOREM Studies/COEAs

Year	Study (S)/COEA (C)/Reimbursable (R)
85	FADEWS (S)
85	STINGER Proficiency (S)
86	Armor Investment Strategy (S)
87	M1A2 (C)
87	FAADS (C)
87	Infantry Anti Armor Weapon System (IAAWS) (C)
88	FAADS FDTE (M-T-M)
88/89	LHX (C)
88/89	Longbow (C)
89/90	WAM (C)
89/90	Armored Systems Modernization (ASM) (C)
90	AMPAW/ARDEC (R)
91	Auto Tracker (S)
91	Legal Mix VII (C)
91	Army Mortar Master Plan (R)
92	STINGRAY (C)
92	FAADS (C)
92	LOSAT Countermeasures (R)
92	Lightweight Laser Designator/Rangefinder (C)
92/93	Battlefield Combat ID System (BCIS) (C)
92/93	AFAS (C)/SADARM (C)
93	Counter Battery vs NLOS (R)
93	JAVELIN IOTE (S)
93	Division Air Defense (S)
93	Second Generation FLIR (C)
93	Guardian Task Force (R)
93	M1A2 (C)
92/93	M1A2 Synthetic Environment Experiment
93	CR-UAV (C)
93	M2A3 (C)
93	155 SADARM (R)
94	NBC Recon System IOTE (R)
94	MLRS Extended Range Guided Round (R)
94	2K Study for EELS (S)
94	ARPA Jamming (R)
94	TUG-V (R)
94/95	AWS-H Quick Reaction Study (S)
94/95	Anti Armor ATD (R)
94/95	M1 Breacher (C)
94	Land Warrior (C)
94	JAVELIN (R)
94	M1A2 IOTE (M-T-M) (S)
94	Anti-Helicopter Mines (R)
94	Engagement Situational Awareness ®

Table II. Past CASTFOREM Studies/COEAs (Continued)

Year	Study (S)/COEA (C)/Reimbursable (R)
94	V22 Navy COEA (R)
94/95	Off Route Smart Mine Clearing (R)
95	Longbow (C)
95	Degraded States (R)
95	Longbow IOTE (M-T-M) (S)
95	Longbow Countermeasures (C)
95/96	Anti Armor Resource Requirements (S)
95/96	Contermine Tactics (R)
95	BCIS DDL (R)
95/96	WAM IOTE (S)
96	Combined Arms Command & Control (R)
96	Task Force XXI (S)
96	Legal Mix VIII (S)
96	Task Extended Range Munition (R)
96	AAAV (R)
96	International Combat ID (S)

Table III. CASTFOREM Comparisons to Other Simulations

CASTFOREM Compared to	Important Insights Gained
1) CARMONETTE	-CASTFOREM would have provided same results if it had been used in Armor Investment Strategy (AIS) Study -End game statistics were comparable -Statistics over battle time were comparable
2) Janus	-Highlighted the differences due to: False Targets Overkill Acquisition Level Required for Trigger Pull Use of Vegetation Bradley Crossover Range: Missile-to-Gun
3) Ground Wars	-Helped determine draw methodology for acquisition using NVL P-Infinity
4) SIMNET-D	-Motivated change in tank gunner disengage logic: "Shoot until he stops" -Commander uses CITV less than 100 percent of time
5) ModSAF	-CASTFOREM has increased fidelity of some missile flyouts -This is ongoing presently

Table IV. CASTFOREM Field Test/Battle Comparisons

	Comparison to Test/Battle	Modeling Insights/Changes
1	Armor Combat Operations Model Support (ARCOMS) Field Test Experiment Phase II at Ft Hood (1986)	-Shots vs range compared well -Engage times same for defender but longer in CASTFOREM for attacker
2	Soviet Artillery Effects (SAE) (1987)	-Attrition trends between personnel and armored vehicles agreed with test but not with truck
3	Smoke Week 5B Clear Air Trials vs NVL Predictions for Probability of Detection (1988)	-Good FIT for FLIR -Poor FIT for OPTICS
4	Forward Area Air Defense Systems Initial Operation Test and Evaluation (IOTE) (1989)	-LOS-F-H and PMS average shot range from model and test were within 1 sigma of each other -Explicit field test movement scripted into model via external events for first time
5	AMSAA Multiple Target Acquisition Study (1990)	-Suggested that each observer has a "detect threshold" -As target size increases, probability of detect increases
6	Study of Artillery Effects (SAE) Phase IIA Technical Shoot (1990)	-Carelton damage function over estimates damage vs truck and underestimates vs armor
7	M1A2 EUTE (1992-93) M1A2 IOTE (1993-94)	-Range-time scatter plots of shots correlated well -IOTEgunners fired conventional rounds at ranges > 3000m
8	JAVELIN IOTE (1993)	-Limited amount of pre-test work done only
9	NBCRS IOTE (1994)	-Used field test to model tactics of encountering a contaminated area
10	Apache Longbow IOTE (1995)	-Cross walk only for COEA linkage -Helicopter sequence of tactics compared well -Timelines were longer in field (still an open issue)
11	WAM IOTE (1995-96)	-Code peelback by OEC -AMSAA review of data
12	Task Force XXI	-Baseline case in progress -Digitization case near future
13	Distributed Interactive Simulation Search and Target Acquisition fidelity (DISTAF)	-Analysis of data in progress to support play of MIS identification
14	73 EASTING-SWA	-P-infinity curve is upper bound to detect ranges -Disabled variable contrasts until more data
15	Norfolk-SWA	-Used for Combat ID sensitivities and accreditation for BCIS COEA

REFERENCES

[1] Neelamkavic, F. (1987), *Computer Simulation and Modeling,* John Wiley.

[2] Law and Kelton, (1991), *Simulation Modeling & Analysis,* 2ed, McGraw-Hill.

[3] AMSAA (1995), *Compendium of High Resolution Attrition Algorithms,* (Second Draft).

[4] CASTFOREM Documentation, *V&V.*

[5] Newman and Walters, *Search and Acquisition in Obscurant Trials,* SMOKE 5-b Test Report, DELNV-TR0043, U.S. Army Communications and Electronics Command (CECOM), Center for Night Vision and Electro-Optics (CCNVE-O), Fort Belvoir, Virginia 22060, November 1984.

[6] S. Tashiro, *Adventures in Comparing Field Test Data to Acquisition Models,* U.S. Army TRAC-WSMR, White Sands Missile Range, New Mexico 88002.

[7] Baker, David, *A Multiple Target Acquisition Test Using Scale Models,* 27 November 1989, AMSAA.

[8] D.W. Hoock, R.A. Sutherland, and D. Clayton, *Combined Obscuration Model for Battlefield Induced Contaminants (COMBIC),* Volume II, EOSAEL 84, October 1984, TR-0160-11.

[9] COL C.G. Kelly, Memorandum for Director TRADOC Analysis Command - WSMR, 24 October 1991, Subject: Enhanced Chemical Representation in CASTFOREM.

[10] D. Hemingway, *Comparisons of ARCOMS II with CASTFOREM, JANUS, CARMONETTE,* Memorandum for Record, 28 May 1986.

[11] Report, IER-OT-1366, *Independent Evaluation of the AVENGER Initial Operational Test and Evaluation,* 30 January 1990 (SECRET).

[12] Gary J. Marchand, *A Study of The Acquisition Process Within CASTFOREM*, TRAC-WSMR, May 1993.

[13] FAADS LOS-R and LOS-F-M M-T-M Study, TRAC-WSMR, 1989.

INTENTIONALLY LEFT BLANK.

EMPIRICAL PROCESSES AND LEAST–SQUARES ESTIMATION

Joseph C. Collins
U. S. Army Research Laboratory
Aberdeen Proving Ground, MD 21005

ABSTRACT

The theory of continuous regression for Gaussian stochastic processes gives rise to a method for computing parametric and nonparametric estimators of the unknown probability density function f of a random sample. The method generalizes to the case in which the observation density is $\mathcal{K}f$, where $\mathcal{K}$ is an arbitrary known operator. Parametric estimators enjoy the usual optimal properties. Nonparametric estimators are obtained by constrained optimization of a quadratic functional in f and are hence easily computable with existing software. Theoretical properties of the estimators, examples with real data, and a simulation study are included.

INTRODUCTION

BACKGROUND

Let the random variables $t_1, \ldots, t_n$, be independent, identically distributed (i.i.d.) with cumulative distribution function (c.d.f.) F_θ. The empirical c.d.f., $F_n(t) = n^{-1} \sum_{i=1}^{n} I(t_i \leq t)$, converges to a Gaussian stochastic process in the sense that

$$\sqrt{n}\,(F_n - F_\theta) \xrightarrow{d} B \circ F_\theta \ \text{ as } \ n \to \infty, \tag{1}$$

where B is a Brownian bridge, which is a zero-mean Gaussian stochastic process with covariance function $E[B(s)B(t)] = s \wedge t - st$. To estimate θ, we model F_n as

$$F_n(t) = F_\theta(t) + n^{-1/2} A_\theta(t), \tag{2}$$

where A_θ is a zero-mean Gaussian process with covariance $E[A_\theta(s)A_\theta(t)] = F_\theta(s \wedge t) - F_\theta(s)F_\theta(t)$.

To provide heuristic motivation for the model, solve (1) for F_n. We may estimate the parameter of model (2) by the methodology of continuous regression for Gaussian stochastic processes, which we now review.

CONTINUOUS REGRESSION

The following development is due to Parzen[1,2,3]. Let $X(t) = M(t) + A(t)$ be a Gaussian stochastic process on a domain $I \subseteq \mathbf{R}$ with unknown mean $E[X(t)] = M(t)$, and known covariance $E[A(s)A(t)] = K(s,t)$. We wish to estimate M by the principle of maximum likelihood. So we need to identify an appropriate likelihood ratio, the definition of which involves some preliminary constructions.

First of all, $L(X(t)) = \{\sum_{i=1}^{n} a_i X(t_i) : n \in \mathbf{N},\ t_i \in I,\ a_i \in \mathbf{R}\}$ is a vector space with inner product $\langle u, v\rangle = E[uv]$ and $\langle \cdot, \cdot \rangle$-completion $L_2(X(t))$, which is a Hilbert space.

Then, H_K is the reproducing kernel Hilbert space (RKHS) with reproducing kernel K and inner product $\langle \cdot, \cdot \rangle_K$. Denote $K(\cdot, t)$ by $K_t(\cdot)$. The fundamental reference is Aronszajn[4].

Next, $\phi_K : H_K \to L_2(X(t))$ is a function characterized by $\phi(K_t) = X(t)$.

Finally, $Y(t)$ zero-mean is a Gaussian process with $E[Y(s)Y(t)] = K(s,t)$, and $P(K)$ and $P(K,M)$ are the probability measures induced by $Y(t)$ and $X(t)$ respectively on a suitable space of sample paths.

With all this in place, we can write down the likelihood ratio. When $M \in H_K$, the measures $P(K)$ and $P(K,M)$ are equivalent, and the Radon-Nikodym derivative of $P(K,M)$ with respect to $P(K)$ is

$$L = \frac{dP(K,M)}{dP(K)}(X) = \exp\left[\phi_K(M) - \tfrac{1}{2}||M||_K^2\right].$$

The maximum likelihood estimator (regression estimator) of $E[X]$ is that value of M which maximizes the likelihood ratio L. It is illuminating to note that in the case of a finite domain I, one obtains the usual least–squares (LS) estimator for M, and in the linear model when $E[X] = M = Z\beta$, the estimator is the familiar weighted linear regression estimator $\hat{\beta} = (Z^T K^{-1} Z)^{-1} Z^T K^{-1} X$.

Sequences of processes $X_n(t) = M(t) + n^{-1/2}A(t)$ have "scaling" properties, which imply that

$$L_n = \frac{dP(K_n,M)}{dP(K)}(X_n) = \exp\left[n\phi_K(X_n,M) - \tfrac{n}{2}||M||_K^2\right] = \left[\frac{dP(K,M)}{dP(K)}(X_n)\right]^n.$$

The form of the likelihood ratio does not depend on sample size. However, our processes have unknown covariance. So we use a modified version of this estimation scheme.

THE PARAMETRIC LEAST–SQUARES ESTIMATION SCHEME

In our models, the mean and covariance share a common parameter $\theta \in \Theta$. The basic model is $X(t) = M_\theta(t) + A_\theta(t)$, with mean $E[X(t)] = M_\theta(t)$ and covariance $E[A_\theta(s)A_\theta(t)] = K_\theta(s,t)$. We obtain a sequence of least–squares estimators for θ. Given θ_0, the sequence $(\theta_1, \theta_2, \theta_3, \ldots)$ is defined by

$$\frac{dP(K_{\theta_i}, M_{\theta_{i+1}})}{dP(K_{\theta_i}, 0)}(X) = \sup\left\{\frac{dP(K_{\theta_i}, M_\theta)}{dP(K_{\theta_i}, 0)}(X) : \theta \in \Theta\right\}.$$

In light of the scaling property, we can apply this concept to any empirical stochastic process X_n, where

$$\sqrt{n}\,(X_n - M_\theta) \xrightarrow{d} A_\theta \ \text{ as } \ n \to \infty,$$

through use of the model

$$X_n(t) = M_\theta(t) + n^{-1/2}A_\theta(t).$$

For a given observation (data process) X_n and initial parameter guess $\theta_{n,0}$, we define the sequence of estimators $(\theta_{n,1}, \theta_{n,2}, \theta_{n,3}, \ldots)$ by

$$\frac{dP(K_{\theta_{n,i}}, M_{\theta_{n,i+1}})}{dP(K_{\theta_{n,i}}, 0)}(X) = \sup\left\{\frac{dP(K_{\theta_{n,i}}, M_\theta)}{dP(K_{\theta_{n,i}}, 0)}(X) : \theta \in \Theta\right\}.$$

We use the likelihood ratio for known covariance in a recursive fashion to estimate the parameter. Note that this scheme is not limited to estimation based on the empirical c.d.f. F_n, but applies to any suitable (asymptotically Gaussian) empirical stochastic process.

DISTRIBUTION OF THE PARAMETRIC LEAST–SQUARES ESTIMATOR

We present the following result concerning the consistency and asymptotic distribution of the general parametric LS estimator without proof. Under suitable regularity conditions, the first-stage ($i = 1$) estimator behaves according to

$$\sqrt{n} \cdot (\theta_{n,1} - \tau) \xrightarrow{d} N\left(0, \frac{||S_{\gamma\tau}\dot{M}_\tau||_\gamma^2}{||\dot{M}_\tau||_\gamma^4}\right) \ \text{ as } \ n \to \infty,$$

and for all $i > 1$, the iterated estimators according to

$$\sqrt{n} \cdot (\theta_{n,i} - \tau) \xrightarrow{d} N\left(0, \frac{1}{||\dot{M}_\tau||_\tau^2}\right) \ \text{ as } \ n \to \infty,$$

where

n is the sample size,
$\{M_\theta : \theta \in \Theta\}$ is a parametric family of mean value functions,
$\dot{M}_\theta(t)$ denotes $\frac{d}{d\theta}M_\theta(t)$,
τ is the true parameter value,
$\gamma = \theta_{n,0}$ is an initial parameter guess,
$\theta_{n,i}$ is the estimator, assuming covariance parameter $\theta_{n,i-1}$, and
$S_{\gamma\tau}$ represents the square root of the map $K_\gamma(t,\cdot) \mapsto K_\tau(t,\cdot)$.

The parametric estimator is asymptotically unbiased and has a normal distribution. Asymptotic distributions are the same for all iterates $i \geq 2$.

Information and Optimality. Fisher's information measure (Rao[5]) is

$$\begin{aligned} I(\theta) &= E_\theta\left[\left(\frac{d}{d\theta}\log\frac{dP_\theta}{dP_0}\right)^2\right] = E_\theta\left[\left(n\phi(X_n,\dot{M}_\theta) - n\left\langle M_\theta,\dot{M}_\theta\right\rangle_\theta\right)^2\right] \\ &= n^2\mathrm{Var}_\theta\left[\phi(X_n,\dot{M}_\theta)\right] = n\|\dot{M}_\theta\|_\theta^2. \end{aligned}$$

For $i > 1$, the variance of the estimator achieves the Cramér-Rao lower bound as $n \to \infty$. Therefore, the iterated LS estimator with $i = 2$ is "efficient," or asymptotically optimal and therefore equivalent to the maximum likelihood estimator (MLE).

DENSITY ESTIMATION

Consider the special case of density estimation based on $X_n = F_n$. Let $t_1,\ldots,t_n$ be i.i.d. with c.d.f. $F_\tau(t)$ and probability density function (p.d.f.) $f_\tau(t)$. The negative log LS functional assumes the form

$$J_{n,\gamma}(\theta) = -\int \frac{f_\theta(t)}{f_\gamma(t)}\,dF_n(t) + \frac{1}{2}\int \frac{f_\theta(t)^2}{f_\gamma(t)}\,dt.$$

The LS estimator sequence is $(\gamma = \tau_{n,0},\ \tau_{n,1},\ \tau_{n,2},\ \ldots)$, where $J_{n,\tau_{n,i-1}}(\tau_{n,i}) = \inf\left\{J_{n,\tau_{n,i-1}}(\theta) :\ \theta \in \Theta\right\}$. It can be shown that for any n, if $\tau_{n,i}$ converges as $i \to \infty$, then $\tau_{n,\infty}$ minimizes

$$\tilde{J}_n(\theta) = -\int \log f_\theta(t)\,dF_n(t),$$

which is the negative log likelihood. Thus, $\tau_{n,\infty}$ is the traditional MLE.

Example: Linear Density. Let $t_1,\ldots,t_n$ be i.i.d. on $[0,1]$ with density $f_\tau(t) = \tau + 2(1-\tau)t$, where $\tau \in [0,2]$. With γ fixed, the LS functional is

$$J_{n,\gamma}(\theta) = -\int_0^1 \frac{\theta + 2(1-\theta)t}{\gamma + 2(1-\gamma)t}dF_n(t) + \frac{1}{2}\int_0^1 \frac{(\theta + 2(1-\theta)t)^2}{\gamma + 2(1-\gamma)t}dt.$$

For $\gamma = 1$, we get $\tau_n = 4 - \frac{6}{n}\sum_{i=1}^n t_i = 4 - 6\bar{t}$. If $\gamma \neq 1$, the LS estimator is

$$\tau_n = \gamma + \frac{(\gamma-1)^3\sum_{i=1}^n \frac{1-2t_i}{\gamma+2(1-\gamma)t_i}}{n\left(1-\gamma-\frac{1}{2}\log\left(\frac{2}{\gamma}-1\right)\right)}.$$

For comparison, the MLE is the solution θ of

$$\sum_{i=1}^n \frac{1-2t_i}{\theta + 2(1-\theta)t_i} = 0,$$

which is not obtainable in closed form.

Simulation: Linear Density. To illustrate these calculations, we conduct a small simulation using the density function of the previous section, $f_\tau(t) = \tau + 2(1-\tau)t$. The true parameter value is $\tau = 0.333$. Three sample sizes are used: $n = 10$, $n = 100$, and $n = 1000$. In all cases, the initial guess for LS estimation is $\gamma = \theta_{n,0} = 1.0$. The simulation is based on $N = 1000$ runs. The mean squared errors (MSE) presented in the body of Table 1 are calculated as

$MSE = \frac{1}{N}\sum_{k=1}^{N}(\tau - \theta_{n,i}(k))^2$, where $\theta_{n,i}(k)$ is the LS estimate for the i^{th} iterate of the k^{th} simulation run using a sample size of n. Likewise for the maximum likelihood estimator $\hat{\theta}$. Even the first-stage estimator $\theta_{n,1}$ has acceptable properties, compared to the MLE $\hat{\theta}$.

As stated, the LS estimation scheme can be applied to other stochastic processes.

POISSON PROCESS INTENSITY ESTIMATION

Here, we consider a Poisson process. Let $N(t) = \sum_{i=1}^{N} I(T_i \le t)$ be the counting process for a Poisson process with intensity g and mean measure $G = \int g$. The model is

$$N(t) = G(t) + A(t),$$

where A is a Gaussian process with mean $E[A(t)] = 0$ and covariance $E[A(s)A(t)] = G(s \wedge t)$. The LS functional is

$$J_\gamma(\theta) = -\int \frac{g_\theta(t)}{g_\gamma(t)}\, dN(t) + \frac{1}{2}\int \frac{g_\theta(t)^2}{g_\gamma(t)}\, dt.$$

The functional has the same form as in density estimation even though the covariance structures are different. The convergent estimator is the MLE in this case also.

QUANTILE FUNCTION

The quantile function Q is the inverse of the c.d.f. F. Likewise, Q_n is the inverse of F_n. Our model for the empirical quantile process is

$$Q_n(u) = Q(u) + n^{-1/2}A(u),$$

where A is a Gaussian process with mean $E[A] = 0$ and covariance $E[A(u)A(v)] = Q'(u)Q'(v)(u \wedge v - uv)$. The LS functional for this process is

$$J_{n,\gamma}(\theta) = -\int_0^1 \left(\frac{Q_\theta(u)}{Q'_\gamma(u)}\right)' \left(\frac{Q_n(u)}{Q'_\gamma(u)}\right)' du + \frac{1}{2}\int_0^1 \left[\left(\frac{Q_\theta(u)}{Q'_\gamma(u)}\right)'\right]^2 du.$$

The asymptotic covariance of the data process, in this case the quantile function, determines the form of the LS functional.

Location and Scale Estimation for the Quantile Function. For location and scale estimation, we have a fixed quantile function Q_o. The parameter is $\theta = (a, b)$, and candidate functions have the form

$$Q(u; a, b) = a + bQ_o(u).$$

For any choice of γ, the LS estimator is

$$\begin{bmatrix} a_n \\ b_n \end{bmatrix} = \begin{bmatrix} \langle 1,1\rangle & \langle 1,Q_o\rangle \\ \langle 1,Q_o\rangle & \langle Q_o,Q_o\rangle \end{bmatrix}^{-1} \begin{bmatrix} \langle 1,Q_n\rangle \\ \langle Q_o,Q_n\rangle \end{bmatrix},$$

where the RKHS inner product is $\langle x,y\rangle = \int_0^1 (x/Q'_o)'(y/Q'_o)'$. This estimator is known to be best linear unbiased (Bennett[6], Parzen[7]). Note that the LS estimator is independent of the covariance parameter.

NONPARAMETRIC ESTIMATION

We return to density estimation. Relaxing restrictions on the candidate density functions gives a nonparametric estimation problem. Let $X_1, \ldots, X_n$ be i.i.d. F_o. Let h be a p.d.f. The natural nonparametric version of the density estimation problem is:

$$\text{minimize } J_{n,h}(f) = -\int \frac{f}{h} dF_n + \frac{1}{2}\int \frac{f^2}{h} \quad \text{subject to } f \in L_2, \ f \geq 0, \text{ and } \int f = 1.$$

This problem has no solution. We can define a sequence of f's with J unbounded below. These f's approach "spikes" at the data values. The solution tends to the empirical point measure, $f_n = F_n' = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$.

PENALIZED DENSITY ESTIMATION

We change the problem by adding a penalty term to the objective functional. This term grows larger as f gets close to f_n.

Let $X_1, \ldots, X_n$ be i.i.d. F_o, with $f_o = F_o'$. Let h be a p.d.f. Let $\mathcal{D}$ be a linear differential operator of order $p \geq 1$ with no constant term, and let $\alpha > 0$. Then the problem

$$\text{minimize } J_{n,h}(f) = -\int \frac{f}{h} dF_n + \frac{1}{2}\int \frac{f^2}{h} + \frac{\alpha}{2}\int \frac{(\mathcal{D}f)^2}{h}$$

$$\text{subject to } f \in \mathcal{H}_p, \ f \geq 0, \text{ and } \int f = 1$$

has a unique solution (by a theorem of Thompson and Tapia[8]).

The spaces $\mathcal{H}_p = \{f : f^{(p)} \in L_2\}$ are Sobolev spaces. The "correct" weight is $h = f_o$. Penalized estimation has been investigated by Good and Gaskins[9], Silverman[10], Thompson and Tapia[11], Wahba[12], Cox[13], O'Sullivan[14], and many others.

Continuous Representation. The LS functional can be expressed in terms of weighted L_2 inner products, so we can write out the differential equation that characterizes the estimator. Inner products are $\langle x, y \rangle = \int xy$ and $\langle x, y \rangle_w = \int xyw$. Identifying $w = 1/h$, the LS functional is

$$\begin{aligned} J(f) &= -\langle f, f_n \rangle_w + \tfrac{1}{2}\|f\|_w^2 + \tfrac{\alpha}{2}\|\mathcal{D}f\|_w^2 = -\langle f, wf_n \rangle + \tfrac{1}{2}\langle f, wf \rangle + \tfrac{\alpha}{2}\langle \mathcal{D}f, w\mathcal{D}f \rangle \\ &= -\langle f, wf_n \rangle + \tfrac{1}{2}\langle f, (w + \alpha\mathcal{D}^* w\mathcal{D})f \rangle . \end{aligned}$$

The estimator satisfies the differential equation $(w + \alpha\mathcal{D}^* w\mathcal{D})f = wf_n$ subject to $f \geq 0$ and $\int f = 1$.

Discrete Representation and Calculation. The discrete version of the LS functional is

$$J(f) = -f_n^T Rf + \tfrac{1}{2} f^T Rf + \tfrac{\alpha}{2} f^T D^* RDf.$$

Equivalently, with $\langle x, y \rangle_R = x^T Ry$, we can take $J(f) = \|f - f_n\|_R^2 + \alpha\|Df\|_R^2$. Then, the LS estimation problem

$$\text{minimize } J(f) \text{ subject to } f \geq 0 \text{ and } \int f = 1$$

is the standard quadratic programming problem. The corresponding matrix equation for f is

$$(R + \alpha D^* RD)f = Rf_n.$$

Quadratic programming problems are easy to solve, in the sense that high-quality software is widely available. All calculations in this report were performed using Visual Numerics, Inc., IMSL routines.

Figures 1 through 4 show the effects of varying the parameters that characterize the LS estimator. The data set used is the "Buffalo Snowfall" data, which is sometimes exhibited as a sample from a trimodal distribution. Figure 1 shows the effect of discretization grid size. The sizes depicted are 10, 20, 50, and 100. Figure 2 shows the effect of changing the smoothing parameter α. Figure 3 shows the effect of iterating the estimator. There are five curves on this graph. For all practical purposes, convergence is complete by the third iteration. Various derivative penalty functionals were used in Figure 4.

CHARACTERIZATION OF THE LEAST-SQUARES ESTIMATOR

By the superposition principle for linear differential equations, the LS density estimator can be written as a sum. The unconstrained solution in the continuous representation is

$$f(t) = \frac{1}{n}\sum_{i=1}^{n} Z_{\alpha,X_i}(t),$$

where $Z_{\alpha,X}$ satisfies the differential equation $(w+\alpha\mathcal{D}^*w\mathcal{D})Z_{\alpha,X} = w\delta_X$. A density estimator of this form is referred to as a generalized kernel density estimator.

In fact, for $\mathcal{D}x = x'$ and $h \equiv$ uniform, the LS differential equation is $Z_{\alpha,X} - \alpha Z''_{\alpha,X} = \delta_X$ and the solution on $[-T,T]$ becomes, as $T \to \infty$, $Z_{\alpha,X}(t) = \frac{1}{2\sqrt{\alpha}}\exp\left(-|X-t|/\sqrt{\alpha}\right)$. Obviously, this is the standard kernel density estimator, with a bilateral exponential kernel.

Consistency. Under some technical assumptions, which will not be enumerated here, a result of Bosq and Lecoutre[15] on generalized kernel density estimators gives strong uniform consistency. Consistency requires a sequence of smoothing parameters α_n that go to zero, but not too quickly.

Let w be fixed. $\mathcal{D}$ is of order p. Let $\hat{f}_n$ be the minimizer of

$$J_{n,\alpha_n,w}(f) = -\langle f, f_n\rangle_w + \tfrac{1}{2}\|f\|_w^2 + \tfrac{\alpha_n}{2}\|\mathcal{D}f\|_w^2.$$

If

$$\alpha_n \to 0 \;\text{ and }\; (n/\log n)^{2p}\alpha_n \to \infty \;\text{ as }\; n \to \infty,$$

then

$$E\left[\sup_t |\hat{f}_n(t) - f_o(t)|\right] \to 0 \;\text{ as }\; n \to \infty$$

and

$$\sup_t |\hat{f}_n(t) - f_o(t)| \xrightarrow{a.s.} 0 \;\text{ as }\; n \to \infty.$$

Rates of Convergence. We can provide two different results about the rate of convergence of the LS density estimator.

(1.) Bosq and Lecoutre[16] also provide convergence rates in the supremum norm. For n large enough, there exists a δ such that for any $\varepsilon > 0$

$$P\left[\sup_t |\hat{f}_n(t) - f_o(t)| \geq \varepsilon\right] \leq 2\exp\left(-n\delta\varepsilon^2\alpha_n^{1/2p}\right).$$

This implies

$$\sup_t |\hat{f}_n(t) - f_o(t)| = O_p\left(n^{-1/2}\alpha_n^{-1/4p}(\log n)^{1/2}\right).$$

(2.) An analysis similar to that of Silverman[17] or Cox and O'Sullivan[18] establishes the following result.

Let $w = 1/f_o$. If $\alpha_n \to 0$, and $n^{2p}\alpha_n \to \infty$, then

$$E\left[\|\hat{f}_n - f_o\|_w\right] = E\left[\int \frac{1}{f_o}|\hat{f}_n - f_o|^2\right]^{1/2} = O(n^{-1/2}\alpha_n^{-1/4p} + \alpha_n^{1-1/4p}).$$

This gives a rate of $O(n^{-1/2+1/8p})$ for $\alpha_n \sim n^{-1/2}$. (This rate may be applicable for f_o bounded above and away from 0 on compact support.)

SMOOTHING PARAMETER SELECTION FOR DENSITY ESTIMATION

Practically speaking, we need an automatic procedure for selecting the smoothing parameter. Cross-validation is suited to least–squares problems and has been applied to spline smoothing and other statistical estimation and regression problems. See Wahba[19].

The discrete representation has unconstrained solution $f = M_\alpha f_n$ where $M_\alpha = (R + \alpha D^* R D)^{-1} R$. The generalized cross–validation (GCV) criterion for selection of the smoothing parameter is

$$\text{minimize} \quad C(\alpha) = \frac{\|(I - M_\alpha) f_n\|_R^2}{[\mathrm{Tr}(I - M_\alpha)]^2}.$$

The GCV score $C(\alpha)$ is an estimate of mean squared error (Härdle[20], Wahba[21]).

See Figure 5 for an example of smoothing parameter selection by GCV. A sample of size 100 was drawn from a normal mixture distribution. The different graphs highlighted with the $\times$'s are LS estimates computed using the indicated values of α. The GCV criterion picks the smoothing parameter which gives the graph in the center of the array, with $\alpha = 0.0022$.

INDIRECT OBSERVATION

Let $Z_1, \ldots, Z_n$ be i.i.d. with unknown c.d.f. F_o and p.d.f. f_o. We observe $X_1, \ldots, X_n$ which have p.d.f. $g_o = \mathcal{K} f_o$, where $\mathcal{K}$ is some known operator. The empirical functions g_n and G_n are based on the X_i. The penalized LS functional for estimation of f_o is

$$J(f) = -\langle g_n, \mathcal{K} f \rangle_w + \tfrac{1}{2} \|\mathcal{K} f\|_w^2 + \tfrac{\alpha}{2} \|\mathcal{D} f\|_w^2,$$

where the "correct" weight is $w = 1/\mathcal{K} f_o$.

Continuous Representation. The LS estimator f satisfies the differential equation

$$[(\mathcal{K}' f)^* w \mathcal{K} + \alpha \mathcal{D}^* w \mathcal{D}] f = (\mathcal{K}' f)^* w g_n.$$

The prime denotes Gateaux differentiation, and the asterisk denotes the Hilbert-adjoint operator. If $\mathcal{K}$ is a linear operator, the equation becomes

$$(\mathcal{K}^* w \mathcal{K} + \alpha \mathcal{D}^* w \mathcal{D}) f = \mathcal{K}^* w g_n.$$

Discrete Representation. For linear $\mathcal{K}$, the discrete LS functional is

$$J(f) = -g_n^T R K f + \tfrac{1}{2} f^T K^* R K f + \tfrac{\alpha}{2} f^T D^* R D f.$$

Equivalently, with $\langle x, y \rangle_R = x^T R y$, we can take

$$J(f) = \|K f - g_n\|_R^2 + \alpha \|D f\|_R^2.$$

The LS estimation problem

$$\text{minimize } J(f) \text{ subject to } f \geq 0 \text{ and } \textstyle\int f = 1$$

is the standard quadratic programming problem. The corresponding matrix equation for f is

$$(K^* R K + \alpha D^* R D) f = K^* R g_n.$$

SMOOTHING PARAMETER SELECTION FOR THE INDIRECT PROBLEM

This is similar to the standard density estimation case. The discrete representation has unconstrained solution $f = M_\alpha g_n$, where $M_\alpha = (K^* RK + \alpha D^* RD)^{-1} K^* R$. The concept of generalized cross–validation can be adapted to the case of indirect observation. The GCV criterion in this case is

$$\text{minimize}\, C(\alpha) = \frac{\|(I - KM_\alpha)g_n\|_R^2}{[\mathrm{Tr}(I - KM_\alpha)]^2}.$$

There is an extra K in the score, because Mg_n is an estimate of f, and KMg_n estimates g. Note that g is the distribution of the (observable) data.

We conclude with two examples of problems which fit into the framework of density estimation from indirect observation, the deconvolution problem and the corpuscle problem.

The Deconvolution Problem. Consider the model

$$X_i = Z_i + W_i, \quad 1 \le i \le n$$

where the Z_i are i.i.d. f (unknown) and the W_i are i.i.d. k (known). We observe the X_i and wish to estimate f. The p.d.f. g of the X_i is the convolution of k and f:

$$g(t) = [\mathcal{K}f](t) = [k * f](t) = \int k(t-x) f(x)\, dx.$$

See Figure 6 for an example of estimation and GCV smoothing parameter selection for the deconvolution problem. A sample of size 250 was drawn from the $10\beta(3,5)$ distribution and contaminated with $N(0,4)$ noise. The short wide distribution is the data density, signal + noise. The narrow distribution is the signal that we wish to recover. The various smoothing parameter values indicated give the different graphs highlighted with $\times$'s. The GCV criterion picks the version in the center of the array, with $\alpha = 0.00037$.

Wicksell's Corpuscle Problem. Spheres with random radii are distributed at random uniformly in a solid medium. The sphere radius p.d.f. is f_o, with support $[0, R_M]$. A slice through the medium gives data which are circles, *i.e.*, sphere – slice intersections. The circle radius p.d.f. g is nonlinear in f_o:

$$g(t) = [\mathcal{K}_o f_o](t) = \frac{t \int_t^{R_M} (x^2 - t^2)^{-1/2} f_o(x)\, dx}{\int_0^{R_M} x f_o(x)\, dx}.$$

Define the function f by $f(t) = f_o(t) / \int_0^{R_M} x f_o(x)\, dx$. Then

$$g(t) = [\mathcal{K}f](t) = t \int_t^{R_M} (x^2 - t^2)^{-1/2} f(x)\, dx$$

is linear in f. We can recover f_o, since $f_o(t) = f(t) / \int_0^{R_M} f(x)\, dx$.

See Figure 7 for an example of estimation and GCV smoothing parameter selection for the corpuscle problem. A sample was drawn from the $\beta(5,2)$ distribution to represent the radii of spheres. This distribution is the taller, skewed solid line. It is the information we want to recover from the sample.

After slicing the spheres randomly, we have 148 circle radii, which are the observable data. Their density is the low, wide curve plotted with the dotted line. The various smoothing parameter values indicated give the different graphs highlighted with $\times$'s. The GCV criterion picks the version in the center of the array, with $\alpha = 0.0014$.

Table 1: Mean Squared Error

	$n = 10$	$n = 100$	$n = 1000$
$\hat{\theta}$	0.213	0.0192	0.00211
$\theta_{n,1}$	0.219	0.0229	0.00250
$\theta_{n,2}$	0.222	0.0193	0.00210
$\theta_{n,3}$	0.212	0.0194	0.00211
$\theta_{n,4}$	0.215	0.0191	0.00211
$\theta_{n,5}$	0.213	0.0193	0.00211
$\theta_{n,6}$	0.214	0.0192	0.00211
$\theta_{n,7}$	0.213	0.0193	0.00211
$\theta_{n,8}$	0.215	0.0192	0.00211
$\theta_{n,9}$	0.213	0.0192	0.00211

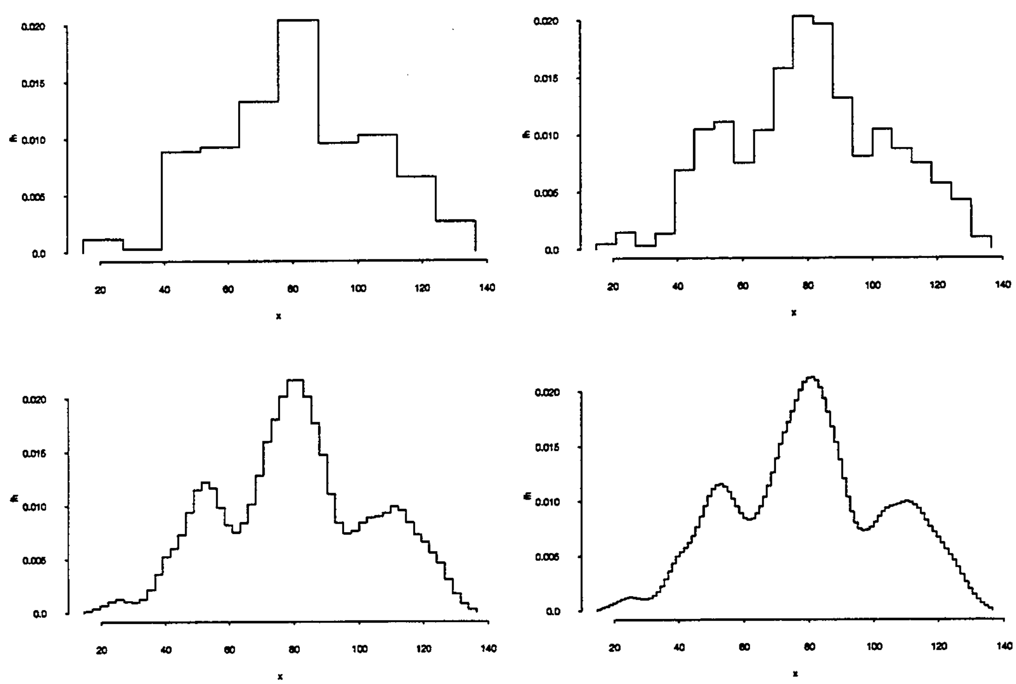


Figure 1: Discretization Effect. Buffalo Snowfall Data, $n = 63$.

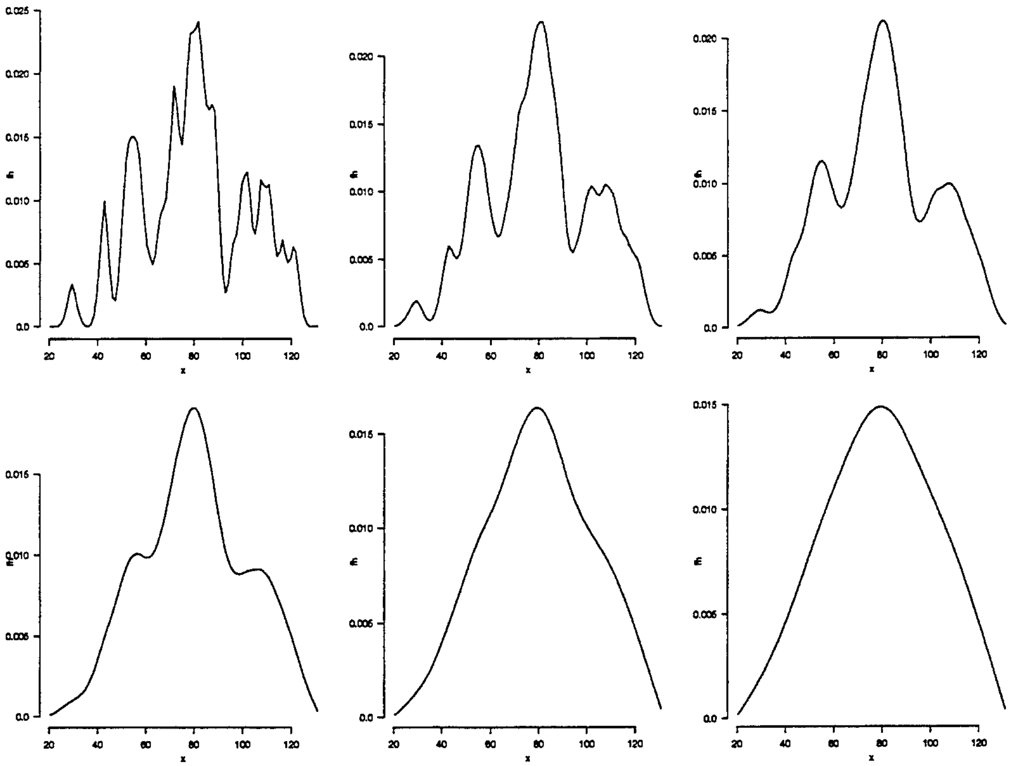


Figure 2: Smoothing Parameter Effect. Buffalo Snowfall Data, $n = 63$.

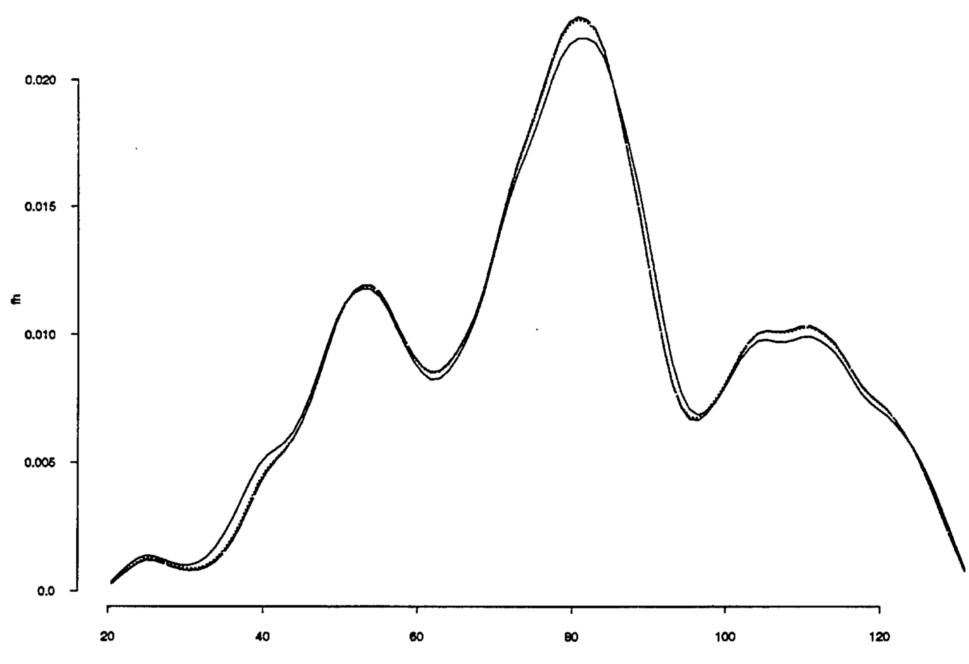


Figure 3: Iteration Effect. Buffalo Snowfall Data, $n = 63$.

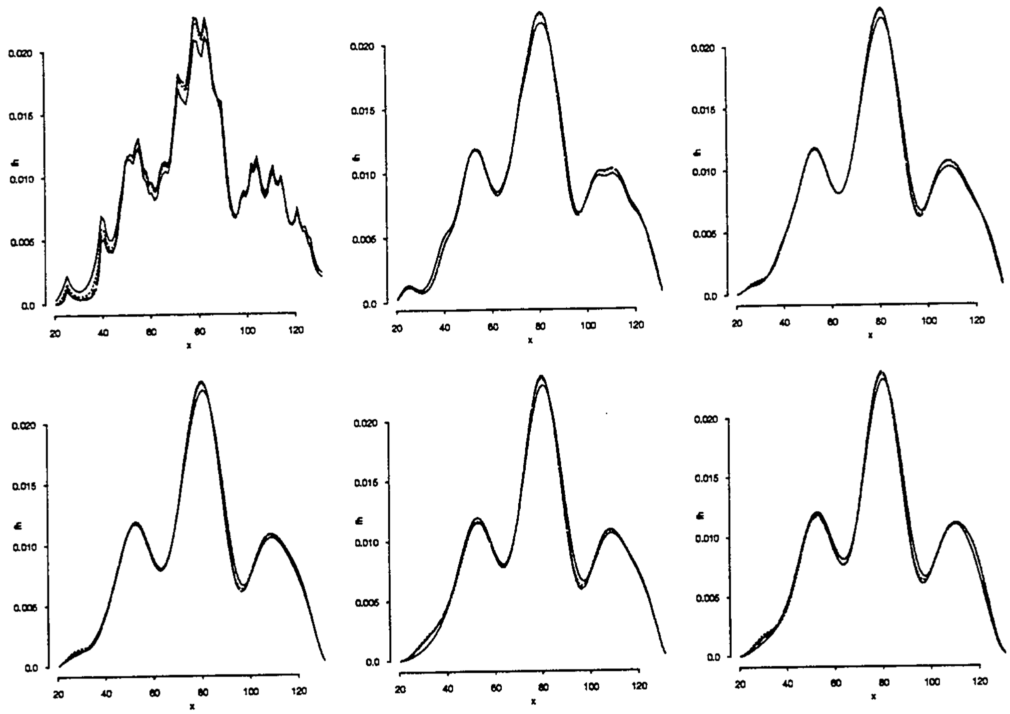


Figure 4: Penalization: $\mathcal{D}x = x^{(p)}$, $1 \le p \le 6$. Buffalo Snowfall Data, $n = 63$.

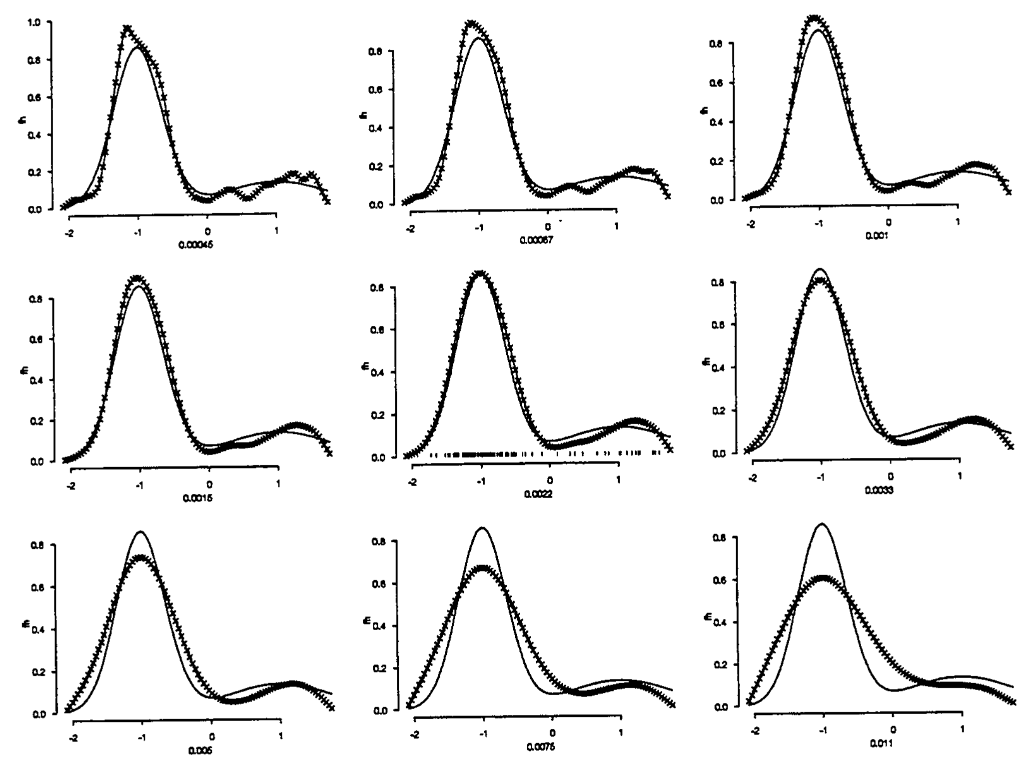


Figure 5: GCV for Density Estimation. $\frac{3}{4}\Phi\left(\frac{x+1}{0.35}\right) + \frac{1}{4}\Phi\left(\frac{x-1}{0.75}\right)$, $n = 100$.

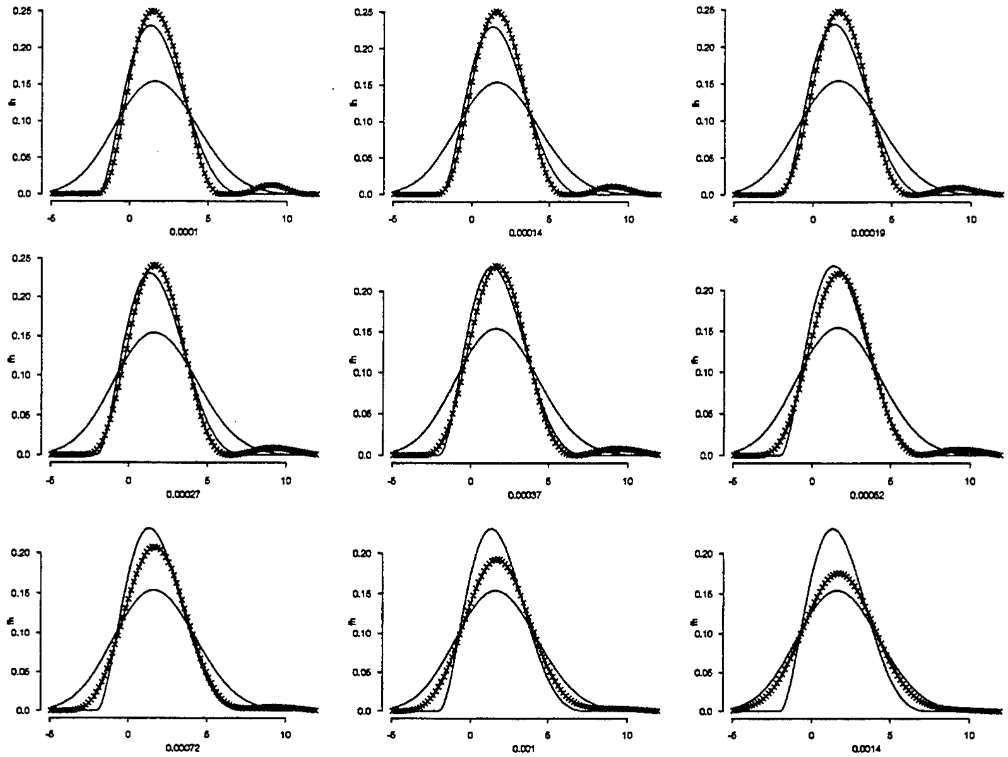


Figure 6: GCV for the Deconvolution Problem. $10\beta(3,5) + N(0,4)$, $n = 250$.

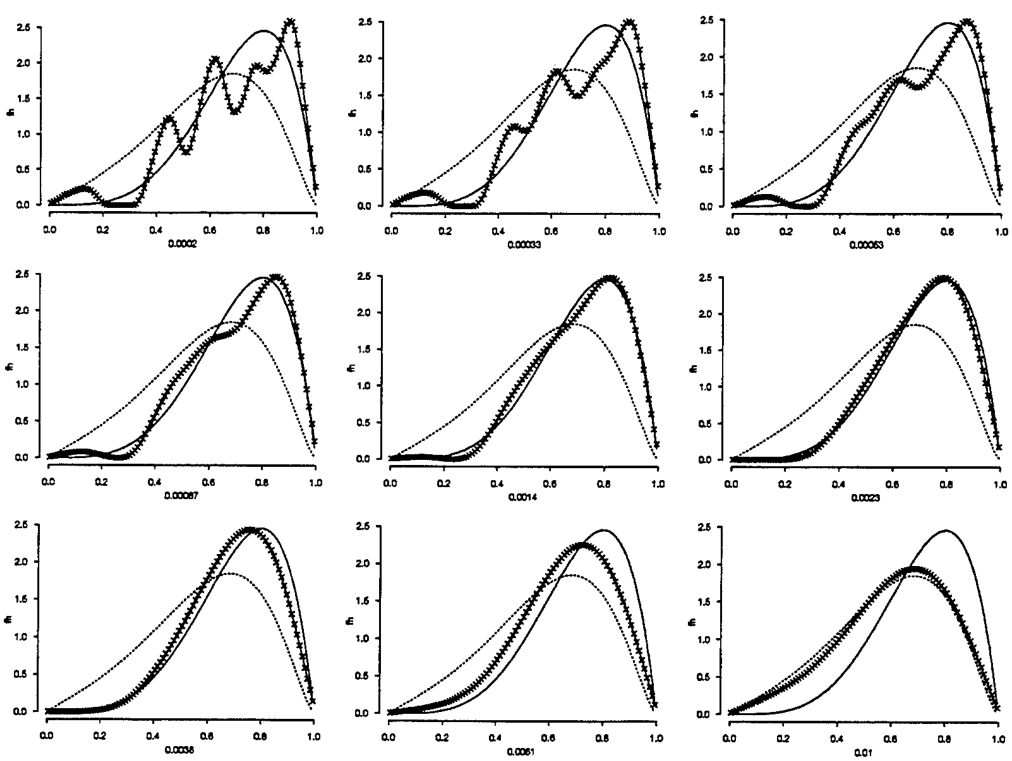


Figure 7: GCV for the Corpuscle Problem. $\beta(5,2)$, $n = 148$.

REFERENCES

1. Parzen, E. "Statistical Inference on Time Series by Hilbert Space Methods, I." Report No. 23, Applied Mathematics and Statistics Laboratory, Stanford University, Stanford, CA, January 1959.

2. Parzen, E. "An Approach to Time Series Analysis." Annals of Mathematical Statistics, vol. 32, no. 4, December 1961.

3. Parzen, E. "Regression Analysis of Continuous Parameter Time Series." Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, p. 469, University of California Press, Berkeley, CA, 1961.

4. Aronszajn, N. "Theory of Reproducing Kernels." Transactions of the American Mathematical Society, vol. 68, p. 337, 1950.

5. Rao, C. R. Linear Statistical Inference and its Applications. New York: John Wiley and Sons, 1973.

6. Bennett, C. A. Asymptotic Properties of Ideal Linear Estimators. Ph.D. Thesis, University of Michigan, 1952.

7. Parzen, E. "Nonparametric Statistical Data Modeling." Journal of the American Statistical Association, vol. 74, no. 365, p. 105, March 1979.

8. Thompson, J., and R. Tapia. Nonparametric Function Estimation, Modeling, and Simulation. Philadelphia: Society for Industrial and Applied Mathematics, 1990.

9. Good, I. J., and R. A. Gaskins. "Nonparametric Roughness Penalties for Probability Densities." Biometrika, vol. 58, no. 2, p. 255, 1971.

10. Silverman, B. "On the estimation of a probability density function by the maximum penalized likelihood method." The Annals of Statistics, vol. 10, no. 3, p. 795, 1982.

11. Thompson and Tapia. *Op. cit.*

12. Wahba, G. Spline Models for Observational Data. Philadelphia: Society for Industrial and Applied Mathematics, 1990.

13. Cox, D. "Approximation of Method of Regularization Estimators." The Annals of Statistics, vol. 18, no. 4, p. 694, 1988.

14. O'Sullivan, F. "A Statistical Perspective on Ill-Posed Inverse Problems." Statistical Science, vol. 1, no. 4, p. 502, 1986.

15. Bosq, D., and J-P Lecoutre. Theorie de l'Estimation Fonctionelle. Paris: Economica, 1987.

16. *Ibid.*

17. Silverman. *Op. cit.*

18. Cox, D., and F. O'Sullivan. "Asymptotic Analysis of Penalized Likelihood and Related Problems." The Annals of Statistics, vol. 18, no. 4, p. 1676, 1990.

19. Wahba. *Op. cit.*

20. Härdle, W. Applied Nonparametric Regression. Cambridge University Press, 1990.

21. Wahba. *Op. cit.*

INTENTIONALLY LEFT BLANK.

PROJECTION METHODS FOR GENERATING MIXED-LEVEL FRACTIONAL FACTORIAL AND SUPERSATURATED DESIGNS

Alonzo Church, Jr.
Church Associates, Inc.
Hudson, Ohio 44236

ABSTRACT

The definitions of resolution and projectivity have been used to develop an algorithm to find mixed-level fractional factorial designs. Some of the designs found differ from standard designs and have superior projection properties. In addition their least squares properties are often superior. The algorithm is discribed and some useful alternative designs are given in detail.

INTRODUCTION

The purpose for this paper is to discuss four subjects related to projection methods: 1) computer generation of mixed-level designs using projectivity criteria; 2) two-level Matryoshka Designs; 3) Selecting projectivity = 3 subsets from published design tables like L36; and 4) projectivity criteria for supersaturated designs. In addition to the above other designs have been generated by the author for certain incomplete latin squares and related designs. In these cases an additional criterion was used for design evaluation. For incomplete latin squares one can use the number of cooccurances of two treatments in the same block and minimize the sums of squares. To show why these designs might prove useful we present the following example.

For a more complete development of projection methods see Church (1993, 1995, 1996)[4,5,6].

PROJECTION GENERATION - AN EXAMPLE

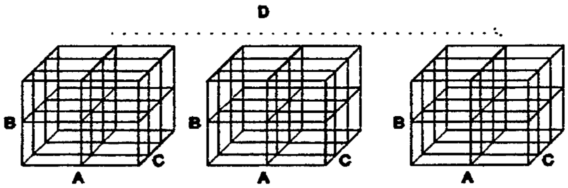
In order to answer questions about the importance of four factors which might affect the wear of a new tennisball design, a production scale experiment was proposed. Two of the factors were to be included at three levels. These factors were discrete settings which could not be reduced to two levels. The other two factors were continuous and two levels were sufficient. The full factorial design would require 36 runs which was too large a number for a production experiment. The factory could tolerate a design requiring 12 or 18 runs but no more. If all factors were discrete the model implied by figure 1 is appropriate. The problem may be visualized as attempting to estimate wear in 36 "boxes" as shown from the 1/3 or 1/2 this number.

Figure 2 shows the EZDOX output listing the runs and design properties for an 18 run experiment. The projection properties of the design are underlined. Included in this listing are some of the design's least squares properties. The output indicates that a typical fractional factorial model requires 7 to 20 parameters (the main effects model requires 7 while the two-factor interaction model requires 20). Thus the full two-factor interaction model cannot be estimated in 18 runs and an analysis appropriate for a supersaturated design seems appropriate.

Some of the least squares properties of the design are included in the output. The design is termed "NON-STANDARD" because it is not orthogonal for the main effects model. Trace efficiencies for the main effects are reported as well as the variance average of the individual contrasts from which these effeciencies are calculated.

The alias index is derived from the alias of main effects due to two-factor interactions. A second output from the EZDOX software is a file (figure 3) which identifies the alias in main effects due to two-factor interactions. The file contains a matrix whose columns are contrasts representing the main effects and whose rows are two-factor interaction contrasts. The calculation used is due to Draper and Smith (1966)[7]. The alias index (AI) is the column sum of squares averaged for factors with more than two levels. The smaller the alias index the less is the bias in main effects due to two-factor interactions.

Four Mixed-Level Discrete Factors
A1,A2 B1,B2 C1,C2,C3 D1,D2,D3



36 Cells in Full Factorial
7 or more Model Parameters
Possible Fractions: 12,18,24,30

Figure 1

Projection Generated Design - Factory Experiment

A	B	C	D	Y	Pred(Y)	Pred(Ye)
1	1	1	2	23.1		
1	1	2	1	21.2		
1	2	1	1	23.9		
1	2	2	2	22.8		
1	3	1	1	23.1		
1	3	2	2	22.3		
2	1	1	1	21.0		
2	1	2	2	18.8		
2	2	1	2	21.8		
2	2	2	1	20.1		
2	3	1	2	21.1		
2	3	2	1	20.8		
3	1	1	1	21.7		
3	1	2	2	19.2		
3	2	1	2	22.8		
3	2	2	1	20.8		
3	3	1	1	21.6		
3	3	2	2	20.8		

Figure 4

Designing a Mixed-Level Experiment

Yesterday ?

Today:

EZDOX Student: 9-26-95 7 to 20 Parameters.
18 Runs, 4 Var, .500 fraction. Levels: 3 3 2 2
Table balance: q2, z2: 1 0 q3, z3: 1 0 q4, z4: 1 1
Design is NON-STANDARD. Resolution = 3. Projectivity = 3.
EFFICIENCIES (%): 99.4 100. 100. 99. 99.
AVERAGE VARIANCE(x10000): 1111 1111 562 562
ALIAS INDEX: 9.3 9.7 9.7 8.9 8.9
C = V1: 1 1 1 1 1 1 2 2 2 2 2 2 3 3 3 3 3 3
D = V2: 1 1 2 2 3 3 1 1 2 2 3 3 1 1 2 2 3 3
A = V3: 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
B = V4: 1 2 2 1 2 1 2 1 1 2 1 2 2 1 1 2 2 1

Figure 2

Non-independent Analyses of Variance
Tentative and Final

	A=?		*B=?*		*C=?*		*D=?*		*Final*	
Source	DF	Adj MS	DF	Adj MS	DF	Adj MS	DF	Adj MS	DF	Adj MS
A	2	6.76	2	6.75	2	6.70	2	6.70	2	6.76
B	2	1.97	2	1.97	2	1.96	2	1.96	2	1.97
C	1	8.70	1	9.71	1	10.30	1	8.27	1	10.58
D	1	.34	1	.00	1	.02	1	.19	-	
A*B	4	.07	4	.07	-		-		-	
A*C	2	.16	-		2	.07	-		-	
A*D	2	.31	-		-		2	.28	-	
B*C	-		2	.88	2	.82	-		2	.98
B*D	-		2	.05	-		2	.05	-	
C*D	-		-		1	.00	1	.00	-	
Error	3	.36	3	.04	6	.06	6	.26	10	.05

Figure 5

Alias Matrix

Alias Matrix = {[INV(X'X)]X'Z}'

2-F.I.	A1	A2	B1	B2	C	D
A1B1	.000	.000	.000	.000	.000	.000
A2B1	.000	.000	.000	.000	.000	.000
A1B2	.000	.000	.000	.000	.000	.000
A2B2	.000	.000	.000	.000	.000	.000
A1C	.000	.000	.000	.000	.000	.000
A2C	.000	.000	.000	.000	-.025	.225
A1D	.000	.000	.000	.000	.000	.000
A2D	.000	.000	.000	.000	.225	-.025
B1C	.000	.000	.000	.000	.000	.000
B2C	.000	.000	.000	.000	-.025	.225
B1D	.000	.000	.000	.000	.000	.000
B2D	.000	.000	.000	.000	.225	-.025
CD	-.222	.444	-.222	.444	.000	.000

Figure 3

Projection Generated Design - Factory Experiment

A	B	C	D	Y	Pred(Y)	Pred(Ye)
1	1	1	2	23.1	23.1	23.1
1	1	2	1	21.2	20.9	20.9
1	2	1	1	23.9	24.1	24.0
1	2	2	2	22.8	22.2	22.2
1	3	1	1	23.1	23.1	23.1
1	3	2	2	22.9	22.5	22.5
2	1	1	1	21.0	21.1	21.1
2	1	2	X	18.8	18.9	18.9
2	2	1	2	21.8	22.0	22.0
2	2	2	1	20.1	20.2	20.2
2	3	1	2	21.1	21.1	21.1
2	3	2	1	20.8	20.4	20.4
3	1	1	1	21.7	21.6	21.6
3	1	2	2	18.2	18.4	18.4
3	2	1	2	22.8	22.5	22.5
3	2	2	1	20.8	20.8	20.8
3	3	1	1	21.6	21.6	21.6
3	3	2	2	20.5	21.0	21.0

Figure 6

Figure 4 lists the designed runs in standard units and the wear response. The purpose of the analysis is to minimize wear. Figure 5 shows the approach used for this design to identify important factors and two-factor interactions. Four tentative analyses were run as indicated with each containing main effects and a different subset of the interactions. The smaller the error mean square the more likely the tentative analysis is to be "correct". Terms in the final model include three main effects and a two-factor interaction. From this model predictions were calculated for each of the 36 "boxes" (figure 6). The minimum wear was identified and after experimental verification used in production.

This experiment was our first application of a design generated by projection methods.

PROJECTIVITY of SOME TWO-LEVEL DESIGNS

In figure 7 we show projections to two and three dimensions of the standard 8 factor 16 run design of resolution IV (see Box and Hunter (1961)[2]). One two-way and one three-way projection are shown. All 56 three-way projections are identical, each of the eight cells has exactly two runs while all 28 two-way projections have cells containing four runs each. Not only is this design of resolution IV but it is also projectivity = 3 by a new criterion proposed by Box and Tyssedal (1992)[3].

If we now consider a 12 factor, 16 run experiment designed by conventional methods, both the resolution and projectivity are reduced as shown in figure 8. We note that the smaller 12 run design shown in figure 9 (due to Plackett and Burman (1946)[13]) has resolution III and projectivity = 3 (Box and Bisgaard (1992)[1]) for just one less factor!

In figures 7, 8, and 9 we have shown projections as if all variables were continuous. At two levels we cannot distinguish in the model between continuous and discrete factors. The model differences become apparent when a factor has three or more levels. It seems more natural to represent discrete factors using tables rather than graphs and identify the factor levels using integers.

We have discussed two criteria due to Box[2,3] and coworkers for classifying designs, resolution and projectivity. We summarize definitions of these properties as follows:

Design Resolution:
- III - Main Effects are aliased with 2 Factor Interactions.
- IV - Main Effects are independent of 2 Factor Interactions.
- V - Main Effects are independent of 2 Factor Interactions and 2 Factor Interactions are independent of one another

Design Projectivity:
- 2 - 2-way tables have no empty cells.
- 3 - 3-way tables have no empty cells.
- 4 - 4-way tables have no empty cells.

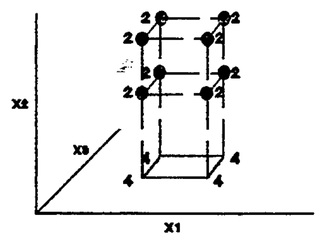
These definitions also extend to higher resolution and projectivity.

In the Box and Tyssedal paper (1992)[3] defining projectivity it is shown that there exists a 16 run design which is projectivity = 3 for up to 14 factors. Using computer search techniques we were able to identify the Box-Tyssedal design as based on a different 16x16 hadamard matrix discussed by Hall (1961)[8]. Hall has given a total of five 16x16 hadamards one of which leads to Taguchi's L16[14] and the usual series of two-level fractional factorials discussed in Box and Hunter[2]. Box and Tyssedal[3] show that three other hadamards lead to projectivity = 3 designs in 12 factors. since the designs are not given in the Box and Tyssedal[3] paper, we present the best of these in the appendix (design 1 and design 2).

We have verified the work of Box and Tyssedal [3] by identifying the best subsets of each of the 5 Hall Hadamards. Our results are summarized in table 1. It should be noted that not all subsets of these hadamards lead to this result. Table 1 tabulates the successful subsets as hits out of the number of possible designs called combinations.

Of the three Hall Hadamards which lead to 12 Factor projectivity = 3 designs one appears best. This design we have named Matryoshka because of its nesting projectivity. In the four a priori most important factors the design is a full factorial. Adding the next four intermediate important factors results in the familiar resolution IV design for 8 factors in 16 runs.

Projection Property - Standard 8-Variable, 16-Run Design
Resolution = IV and Projectivity = 3

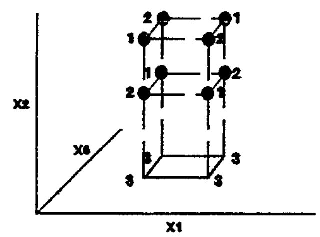


X1	X2	X3	Count
-1	-1	-1	2
1	-1	-1	2
-1	1	-1	2
1	1	-1	2
-1	-1	1	2
1	-1	1	2
-1	1	1	2
1	1	1	2

3-Way Projections: 2 points occupy same space (56 combinations)
2-Way Projections: 4 points occupy same space (28 combinations)

Figure 7

Typical Projection - Standard 11-Variable, 12-Run Design
Resolution = III and Projectivity = 3

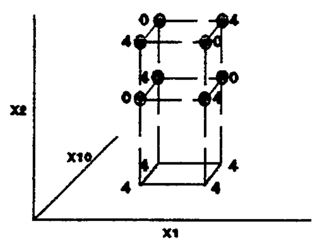


X1	X2	X3	Count
-1	-1	-1	2
1	-1	-1	1
-1	1	-1	1
1	1	-1	2
-1	-1	1	1
1	-1	1	2
-1	1	1	2
1	1	1	1

3-Way Projections: 1 or 2 points occupy same space (165 combinations)
2-Way Projections: 3 points occupy same space (55 combinations)

Figure 9

Projection Property - 12-Variable, 16-Run Design
Resolution = III and Projectivity = 2



X1	X2	X10	Count
-1	-1	-1	0
1	-1	-1	4
-1	1	-1	4
1	1	-1	0
-1	-1	1	4
1	-1	1	0
-1	1	1	0
1	1	1	4

3-Way Projections: Different Types, Some Incomplete!
2-Way Projections: 4 points occupy same space

Figure 8

EZDOX DESIGN ALGORITHM

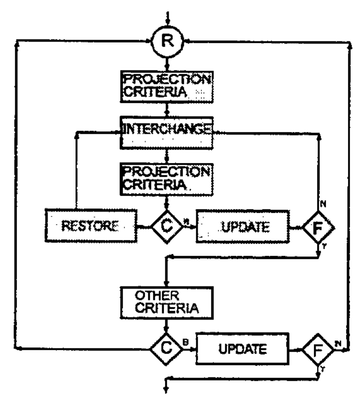


Figure 10

Matroshka permits a bonus however: four more factors of lesser importance can be added! In these twelve factors every three-way table has at least one run per cell. Thus Matryoshka is projectivity = 3. In addition the three remaining degrees of freedom in the design can be used to estimate two-factor interaction groups one containing AB a second containing AC and a third containing BC. The Matryoshka design is given in the Appendix as Design 1.

Table 1
Summary of Designs which are Subsets of the 5 Hall-Hadamard-Based Designs

No. Fac	Comb.	Hall Ref	Hits	z_3	z_4	q_3	q_4	AI	R	P	
14	15	Std	7	28	385	896	6160	.26	III	2	
		2	1	12	593	896	6160	.26	III	2	
		3	3	4	697	896	6160	.26	III	2	
		4	1	0	749	896	6160	.26	III	3	BEST
		5	1	4	749	896	6160	.26	III	2	
12	455	Std	35	16	183	512	2928	.25	III	2	
		2	1	0	327	512	2928	.25	III	3	BEST
		3	3	0	343	512	2928	.25	III	3	
		4	7	0	351	512	2928	.25	III	3	
		5	7	0	351	512	2928	.25	III	3	
8	6435	Std	15	0	14	0	224	.00	IV	3	
		2	3	0	14	0	224	.00	IV	3	
		3	1	0	14	0	224	.00	IV	3	
		4	14	0	34	64	320	.16	III	3	WORST
		5	1	0	14	0	224	.00	IV	3	

Table 2
Number of 2- 3- and 4-way Tables for Selected Number of Factors

Factors	2-Way	3-Way	4-Way
7	21	35	35
11	55	165	330
12	66	220	495
14	91	364	1001
15	105	455	1365
23	253	1771	8855
31	465	4495	31465
47	1081	16215	178365
63	1953	39711	595665
66	2145	45760	720720

Matryoshka can be extended to a larger 32 run design for which 24 factors can be included in a projectivity = 3 design.

It should be noted that for 5 factors the standard design has the best projection properties. It is resolution V and projection = 4.

DESIGN GENERATION vs SUBSET SELECTION

In the preceeding two sections we have demonstrated generating a mixed level design and subsetting Hadamard based designs with good projection properties . These can be viewed as two types of subsetting. If we have an array with columns identified with factors and rows identified with runs then column subsetting can be used when the number of factors is less than the number of columns. It is practical to evaluate all possible subsets to select the best design by the selected criteria.

Row subsetting is usually less practical because of the number of rows and thus combinations to be evaluated. Thus the algorithm used in EZDOX, a row subsetting program, is a directed search from a random start. The directed search proceeds by a single factor level interchange between two rows. The better "design" is kept for the next iteration.

Our criteria which can be used either in row or column subsetting algorithms are as follows:

z_i - number of i-way tables having at least one empty cell.
q_i - Adjusted sum of squared cell counts over all i-way table cells.

The current row subsetting program, EZDOX, uses q_2, q_3, q_4, z_2, z_3, and z_4, to determine search improvement. The minimum

z_i and q_i found define the best design. The q_i relate to design resolution while the z_i relate to design projectivity.

A significant number of i-way tables require evaluation in most practical problems. Table 2 lists the number as a function of factors. While the number of tables does not depend on the number of factor levels the number of cells does.

SEARCH METHOD

Figure 9 is a simplified flowchart for the projection search algorithm used. This algorithm belongs to the class of interchange algorithms. For another example of an interchange algorithm see Nguyen (1996)[12].

The projection algorithm used to generate fractional factorial and supersaturated designs consists of an inner iteration and an outer iteration. The inner iteration uses exclusively the projection criteria. The outer iteration adds criteria appropriate to the design type.

INNER ITERATION

The process begins with a random starting design subject to the constraints that the design size is fixed and the number of factors is fixed as well as the number of levels per factor. It is also a constraint that each level of a factor occur equally often. For this start the projection criteria are calculated.

Next an interchange between two rows of a randomly chosen factor is performed subject to the constraint that the two rows differ in level of the chosen factor. For this interchange the projection criteria are calculated. Should the projection criteria be better than the existing design, the interchange rows replace the original rows.

This interchange process is repeated a large number of times insuring that all row pairs and all factors are included in the interchanges multiple times.

OUTER ITERATION

The outer iteration compares the result of the inner iteration with previous best inner iterations and retains the best of the best. In the outer iteration projection criteria are used in addition to other suitable criteria. For supersaturated designs the other criteria include the maximum $|r|$ among the factors. Also included are average $|r|$ and Det$|RR'|$. For other fractional factorial designs maximum $|r|$ between main effects and two-factor interactions are included. Also included is alias index.

DESIGN GENERATION and COMPARISONS

In tables 3 and 4 we present a comparison of some projection generated designs with the accepted standard designs where one is known. As the accepted standard designs we have used the two-level designs of Box and Hunter (1961)[2], the orthogonal arrays used by Taguchi (1987)[14] and designs used in the MINITAB software. Table 3 contains the comparisons and table 4 contains some designs for which no standard was available to the author. The reference column of the tables indicates where the design details can be found. If no reference is given the details are in the Author's database. This database was generated using EZDOX software. The complete database contains orthogonal designs with factors having 2 to 16 levels and up to 36 runs. Within the limits of the software the database is complete for up to seven factors. Beyond seven factors some scattered designs are included.

COLUMN SUBSETS of L36

Taguchi (1987)[14] has proposed the use of 36 run designs in two and/or three level factors based on a saturated orthogonal array which can accommodate up to 11 two-level and 12 three-level factors. We used this design to determine what the possibilities are for projectivity = 3 designs when six or fewer three-level factors are to be combined with two-level factors in an experiment. We further ask if a better alternative exists.

Table 3
A Comparison of Designs
Accepted Standard with Projection Generated Designs

N	Factor Levels	Type	q_3	z_3	q_4	z_4	AI	Ref
16	12@2	Std	512	16	2928	183	.250	
		PGD	512	0	2928	327	.250	1
16	14@2	Std	896	28	6160	385	.260	
		PGD	896	0	6160	749	.260	2
16	4@2 1@4	Std	16	1	1	4	.149	
		PGD	16	2	1	4	.149	3
36	11@2 1@3	Std	2641	0	13861	330	.244	
		PGD	1	0	3301	0	.081	4
36	8@2 3@3	Std	1126	0	4411	168	.169	
		PGD	661	0	2863	213	.132	6

Table 4
Additional Projection Generated Designs

N	Factor Levels	q_3	z_3	q_4	z_4	AI	Ref
16	4@2 1@8	1	6	1	4	.230	
20	6@2 1@5	81	10	63	35	.186	
24	6@2 1@6	104	5	33	22	.142	
24	6@2 1@12	1	15	1	20	.188	
36	8@2 3@3	661	0	2863	213	.132	
36	6@2 4@3	605	0	2225	172	.131	
36	11@2 2@3	917	0	5493	283	.137	5

Table 5 shows the scope of an exhaustive subsetting of L36. Shown are the number of optimal occurances (hits) and the number of combinations requiring evaluation. Clearly optimal is a rare event except in the case of one three-level factor. We have defined the optimal occurance to be a projection = 3 design with best levels of the other projection properties. Not only did we calculate the best L36 subsets but we also conducted a search to see if designs better than the L36 subsets could be found. In every case studied a better design was found. These better designs have been included in tables 2 and 3.

Table 5
Best Subsets of Design L36
by Projection Criteria

Factor	Levels	Hits	Combinations
11@2	1@3	12	12
10@2	2@3	6	726
8@2	3@3	4	36300
6@2	4@3	1	228690
3@2	5@3	1	130680

Table 6
Small Mixed-Level Supertsaturated PGDs with $|r|<.34$

No.	Factors			2-way			3-way		Reference
Runs	4's	3's	2's	z2	z22	q2	z3	q3	Design
6	0	1	4	0	0	1	10	1	7
12	0	1	18	0	85	340	562	2925	8
	1	0	18	0	76	305	558	2491	9
	1	1	9	0	9	37	71	289	10
18	0	1	29	0	96	673	682	18555	11
	0	2	22	0	59	355	431	8075	12

MIXED-LEVEL SUPERSATURATED DESIGNS

A supersaturated design is a screening design. It is appropriate when a small number of the proposed factors are active. A good rule is that this small number should be less than half the design size. In such a design situation the need for more than two levels in a factor can occur when the factors are discrete. It is not unreasonable to include a small number of such more than two-level factors in a design.

Lin (1993 and 1995)[10,11] gives construction methods for some two-level supersaturated designs and proposes a criterion for useful supersaturated designs. He suggests that no two columns of such designs should have correlation greater in absolute value than 0.34. To compute the correlation among factors in the design it is necessary to model the factors which have more than two levels in such a way that all degrees of freedom are included. For this purpose we have used the orthogonal contrasts.

Modification of EZDOX was required to obtain the supersaturated designs presented here. Only z_2, q_2, z_3, q_3 were used. However it was found necessary to define a supplimental criterion to account for correlation. This new criterion is a measure

of imbalance of two-way tables formed from all pairs of factors. It is defined as:

z_{22} - number of unbalanced 2-way tables not having a zero cell

Tables which cannot be balanced are counted when the imbalance exceeds the best which can be expected.

In this feasibility study only small n with one or two factors at three and/or four levels were studied.A summary of the designs found is given in table 6.

A listing of each design is included in the appendix.

ACKNOWLEDGEMENT

Since the beginning of 1995 the work on projectivity and related matters has been sponsored by Church Associates Inc. primarily for the benefit of clients. Nonclients of Church Associates may obtain written and computer materials from Church Associates but there is a nominal charge.

REFERENCES

1. Box, G.E.P. and S. Bisgaard (1992), "What Can You Find Out from 12 Experimental Runs?" Report 88, CENTER FOR QUALITY AND PRODUCTIVITY IMPROVEMENT, University of Wisconsin-Madison

2. Box, G.E.P. and J. S. Hunter (1961), "The 2^{k-p} Fractional Factorial Designs," TECHNOMETRICS, 3, pp. 311-351 & 449-458

3. Box, G.E.P. and J. Tyssedal (1994), "Projective Properties of Certain Orthogonal Arrays," Report 116, CENTER FOR QUALITY AND PRODUCTIVITY IMPROVEMENT, University of Wisconsin-Madison

4. Church, Alonzo, Jr. (1993), "Projection Methods for Generating Design Arrays," Poster Session Presentation, Gordon Research Conference on Statistics in Chemistry and Chemical Engineering

5. Church, Alonzo, Jr. (1995), "Projection Methods for Generating Mixed-Level Fractional Factorial Designs", ASA Proceedings, Section on Physical & Engineering Sciences

6. Church, Alonzo, Jr. (1996), "Projection Methods for Generating Designs", ASA Proceedings, Section on Physical & Engineering Sciences

7. Draper, N. R. and H. Smith (1966), Applied Regression Analysis, John Wiley & Sons, New York, N. Y.

8. Hall, M.J. (1961), "Hadamard Matrices of Order 16," JET PROPULSION LABORATORY, Summary 1 pp. 21-26

9. Lin, D. K. J. and N. L. Draper (1992),"Projection Properties of Plackett and Burman Designs," Technometrics, 34, pp. 423-428

10. Lin, D. K. J. (1993), "A New Class of Supersaturated Designs", TECHNOMETRICS, 35, pp. 28-31

11. Lin, D. K. J. (1995), "Generating Systematic Supersaturated Designs", TECHNOMETRICS, 37, pp. 213-225

12. Nguyen, N.-K. (1996), "An Algorithmic Approach to Constructing Supersaturated Designs," TECHNOMETRICS, 38, pp. 69-73

13. Plackett, R. L. and J.P. Burman (1946), "The Design of Optimum Multifactorial Experiments," BIOMETRIKA, 33, pp.305-325.

14. Taguchi, G. (1987), System of Experimental Design White Plains, NY: UNIPUB

APPENDIX
Projection Generated Designs Discussed in the Text

Projection 3 or 4 Orthogonal Designs which are better than STANDARD Designs:

Design 1 - 16 Runs, 12 Factors at 2 Levels,
Projection = 3:

F 1: 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2
F 2: 1 1 1 1 2 2 2 2 1 1 1 1 2 2 2 2
F 3: 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2
F 4: 1 1 2 2 2 2 1 1 2 2 1 1 1 1 2 2
F 5: 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
F 6: 1 2 1 2 2 1 2 1 2 1 2 1 1 2 1 2
F 7: 1 2 2 1 1 2 2 1 2 1 1 2 2 1 1 2
F 8: 1 2 2 1 2 1 1 2 1 2 2 1 2 1 1 2
F 9: 2 1 2 1 2 1 1 2 2 1 1 2 1 2 1 2
F10: 2 1 2 1 1 2 2 1 1 2 2 1 1 2 1 2
F11: 2 1 1 2 2 1 2 1 1 2 1 2 2 1 1 2
F12: 2 1 1 2 1 2 1 2 2 1 2 1 2 1 1 2

Design 2 - 16 Runs, 14 Factors at 2 Levels,
Projection = 3:

F 1: 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2
F 2: 1 1 1 1 2 2 2 2 1 1 1 1 2 2 2 2
F 3: 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2
F 4: 1 1 2 2 2 2 1 1 2 2 1 1 1 1 2 2
F 5: 2 2 1 1 1 2 1 2 1 2 1 2 1 1 2 2
F 6: 2 2 1 1 2 1 2 1 2 1 2 1 1 1 2 2
F 7: 1 2 1 2 2 2 1 1 1 1 2 2 1 2 1 2
F 8: 1 2 1 2 1 1 2 2 2 2 1 1 1 2 1 2
F 9: 2 1 2 1 1 2 2 1 2 1 1 2 1 2 1 2
F10: 2 1 2 1 2 1 1 2 1 2 2 1 1 2 1 2
F11: 2 1 1 2 2 1 2 1 1 2 1 2 2 1 1 2
F12: 2 1 1 2 1 2 1 2 2 1 2 1 2 1 1 2
F13: 1 2 2 1 2 1 1 2 2 1 1 2 2 1 1 2
F14: 1 2 2 1 1 2 2 1 1 2 2 1 2 1 1 2

Design 3 - 16 Runs, 5 Factors 1 @ 4 Levels, 4 @ 2 Levels,
Projection = 3:

F 1: 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4
F 2: 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2
F 3: 1 1 2 2 1 2 1 2 1 2 1 2 2 2 1 1
F 4: 1 2 1 2 2 1 1 2 1 2 2 1 1 2 1 2
F 5: 2 1 1 2 2 2 1 1 1 1 2 2 1 2 2 1

Design 4 - 36 Runs, 12 Factors, 1 @ 3 Levels, 11 @ 2 Levels,
Projection = 4:

F 1: 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3
F 2: 1 1 1 1 1 1 2 2 2 2 2 2 1 1 1 1 1 1 2 2 2 2 2 2 1 1 1 1 1 1 2 2 2 2 2 2
F 3: 1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2
F 4: 1 1 2 1 2 2 1 2 2 1 1 2 1 2 2 1 1 2 1 1 2 1 2 2 1 1 2 1 2 2 1 2 2 1 1 2
F 5: 1 2 1 2 1 2 1 2 2 1 2 1 2 1 2 1 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 2 1 2 1
F 6: 1 1 2 2 2 1 2 1 2 1 2 1 2 1 2 1 2 1 1 1 2 2 1 1 2 2 2 1 2 1 2 1 2 1 2 1
F 7: 1 2 2 1 2 1 1 2 1 2 2 1 2 1 1 2 1 2 2 1 2 1 1 2 1 2 2 1 2 1 1 2 1 2 2 1
F 8: 1 2 1 2 2 1 2 2 1 1 1 2 1 2 2 2 1 1 2 1 1 2 1 2 1 2 1 2 2 1 2 2 1 1 1 2
F 9: 1 1 2 2 1 2 2 2 1 2 1 1 2 2 1 2 1 1 1 2 1 1 2 2 1 1 2 2 1 2 2 2 1 2 1 1
F10: 1 2 2 2 1 1 1 1 2 2 1 2 1 2 1 1 2 2 2 2 1 1 1 2 1 2 2 2 1 1 1 1 2 2 1 2
F11: 1 2 2 1 1 2 2 1 1 1 2 2 1 1 2 2 2 1 1 2 2 1 1 2 1 2 2 1 1 2 2 1 1 1 2 2
F12: 1 2 1 1 2 2 2 1 2 2 1 1 2 2 1 1 2 1 1 1 2 2 1 2 1 2 1 1 2 2 2 1 2 2 1 1

Design 5 - 36 Runs, 13 Factors, 2 @ 3 Levels, 11 @ 2 Levels,
Projection = 3:

F 1: 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3
F 2: 1 1 1 1 2 2 2 2 3 3 3 3 1 1 1 1 2 2 2 2 3 3 3 3 1 1 1 1 2 2 2 2 3 3 3 3
F 3: 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2
F 4: 1 2 1 2 1 2 2 2 1 2 1 1 1 2 1 2 1 2 1 1 1 2 2 2 1 2 1 2 1 2 1 2 1 2 1 2
F 5: 1 2 2 1 1 2 1 2 2 1 1 2 2 1 1 1 1 2 2 2 1 2 1 2 1 2 2 2 2 1 1 1 2 1 1 2
F 6: 1 2 2 2 2 2 1 1 1 1 2 1 1 2 1 1 1 1 2 2 2 2 2 1 2 1 2 1 1 2 1 2 2 1 1 2
F 7: 1 2 2 1 2 1 2 1 1 2 1 2 1 1 2 2 1 2 1 2 2 2 1 1 2 2 1 1 2 1 1 2 1 1 2 2
F 8: 2 2 2 1 1 1 1 2 1 2 2 1 1 2 2 1 2 1 1 2 1 2 1 2 1 1 1 2 2 2 1 2 2 1 2 1
F 9: 2 1 2 2 2 1 2 1 2 1 1 1 1 1 2 1 2 2 2 1 1 2 2 1 1 2 1 2 1 2 1 1 2 1 2 2
F10: 2 1 1 1 1 2 2 1 2 1 2 2 1 2 2 2 1 1 2 2 1 2 1 1 2 1 1 2 2 2 1 1 1 2 1 2
F11: 2 2 1 2 2 1 1 2 1 1 1 2 1 2 2 1 1 2 2 1 2 1 1 2 1 1 2 1 2 2 1 1 1 2 2 2
F12: 1 2 1 2 1 2 2 2 1 1 1 2 2 1 2 1 2 1 2 1 2 2 1 1 2 1 1 2 1 1 1 2 2 2 2 1
F13: 2 2 1 1 1 1 2 2 1 2 2 1 1 1 1 2 2 2 2 1 2 1 2 1 2 1 2 2 2 1 1 1 2 1 1 2

Design 6 - 36 Runs, 11 Factors, 3 @ 3 Levels, 8 @ 2 Levels,
Projection = 3:

F 1: 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3
F 2: 1 1 1 1 2 2 2 2 3 3 3 3 1 1 1 1 2 2 2 2 3 3 3 3 1 1 1 1 2 2 2 2 3 3 3 3
F 3: 1 1 2 3 1 2 2 3 1 2 3 3 1 2 3 3 1 1 2 3 1 2 2 3 1 2 2 3 1 2 3 3 1 1 2 3
F 4: 1 2 1 2 2 1 2 1 1 2 1 2 2 2 1 1 1 2 1 2 2 1 2 1 1 1 2 2 1 2 1 2 1 2 1 2
F 5: 1 2 2 1 1 2 2 1 2 1 1 2 1 2 1 2 2 2 1 1 1 1 2 2 1 2 1 2 1 1 2 2 2 2 1 1
F 6: 1 1 2 1 2 2 1 2 1 2 1 2 1 2 2 2 2 1 1 2 2 1 1 1 2 1 2 1 1 1 1 2 2 2 2 1
F 7: 1 2 1 2 1 2 1 1 1 2 2 2 1 2 2 1 1 2 2 1 2 1 1 2 2 2 1 1 2 2 1 2 2 1 1 1
F 8: 2 1 1 2 2 2 1 1 2 1 2 1 1 1 1 2 1 2 2 2 2 1 2 1 1 2 2 2 1 1 1 2 2 1 2 1
F 9: 1 1 2 1 2 1 2 1 2 2 2 1 2 2 1 2 1 2 1 2 1 1 1 2 2 2 1 1 2 1 1 2 1 1 2 2
F10: 2 1 1 2 2 2 2 1 1 1 1 2 1 2 2 2 1 2 1 1 2 2 1 1 1 2 1 1 2 1 2 1 1 2 2 2
F11: 2 2 1 1 1 2 2 2 1 1 2 1 1 2 1 2 1 1 1 2 2 1 2 2 2 1 1 2 2 2 1 1 1 2 2 1

Supersaturated Designs with correlation among factors less than 1/3:

Design 7 - 6 Runs, 5 Factors, 1 @ 3 Levels, 4 @ 2 Levels:
F 1: 1 1 2 2 3 3
F 2: 1 2 1 2 1 2
F 3: 2 1 2 1 1 2
F 4: 1 2 2 1 1 2
F 5: 2 1 1 2 1 2

Design 8 - 12 Runs, 19 Factors, 1 @ 3 Levels, 18 @ 2 Levels:
F 1: 3 3 1 1 3 3 1 2 2 2 2 1
F 2: 2 2 2 2 1 1 1 1 2 1 2 1
F 3: 2 1 2 2 2 1 1 1 1 2 2 1
F 4: 2 1 1 2 2 1 2 1 2 1 2 1
F 5: 2 1 1 2 1 2 2 1 2 2 1 1
F 6: 1 2 1 2 1 2 2 1 1 2 2 1
F 7: 1 2 2 2 2 1 1 2 2 1 1 1
F 8: 1 1 2 2 2 2 1 2 1 1 2 1
F 9: 1 2 1 2 2 1 2 2 1 1 2 1
F10: 1 2 2 2 1 2 1 1 2 2 1 1
F11: 2 2 2 1 1 1 1 1 1 2 2 2
F12: 2 2 2 1 1 1 2 2 2 1 1 1
F13: 1 1 2 1 2 2 1 1 2 1 2 2
F14: 1 2 1 2 1 2 1 1 2 1 2 2
F15: 2 2 1 2 1 1 1 2 1 1 2 2
F16: 1 2 1 1 2 1 2 1 2 1 2 2
F17: 2 1 1 2 1 2 1 2 2 1 1 2
F18: 2 1 2 1 1 2 2 2 1 1 2 1
F19: 2 1 1 2 2 1 1 2 1 2 1 2

Design 9 - 12 Runs, 19 Factors, 1 @ 4 Levels, 18 @ 2 Levels:
F 1: 3 2 4 3 4 2 2 1 3 1 1 4
F 2: 1 2 2 2 1 1 1 1 2 1 2 2
F 3: 2 1 2 2 1 2 1 1 1 2 1 2
F 4: 2 2 1 1 1 2 1 2 1 1 2 2
F 5: 2 2 1 1 1 1 2 1 2 2 1 2
F 6: 2 2 2 1 2 1 1 1 2 1 2 1
F 7: 2 2 1 2 2 2 1 1 1 1 2 1
F 8: 1 2 2 2 1 1 2 2 1 1 1 2
F 9: 2 1 1 1 2 2 1 2 2 1 1 2
F10: 2 1 2 1 1 2 2 1 1 1 2 2
F11: 2 2 2 1 1 1 1 1 1 2 2 2
F12: 2 2 2 1 1 1 2 2 2 1 1 1
F13: 1 1 2 1 1 2 1 2 2 1 2 2
F14: 1 2 1 1 1 2 1 1 2 2 2 2
F15: 2 2 1 1 2 1 1 2 1 2 1 2
F16: 2 1 2 1 1 2 1 2 2 2 1 1
F17: 2 1 1 2 1 1 2 2 1 1 2 2
F18: 2 1 2 1 2 1 2 1 1 2 2 1
F19: 2 2 2 2 1 2 1 2 1 1 1 1

Design 10 - 12 Runs, 11 Factors,
1 @ 4 Levels, 1 @ 3 Levels, 9 @ 2 Levels:
F 1: 3 2 1 3 1 2 2 1 3 1 2 3
F 2: 2 3 1 1 2 4 1 3 4 4 2 3
F 3: 2 2 2 1 1 1 1 1 1 2 2 2
F 4: 2 2 2 1 1 1 2 2 2 1 1 1
F 5: 1 2 2 1 2 2 1 1 2 1 1 2
F 6: 2 1 2 1 1 2 1 2 2 1 2 1
F 7: 1 1 1 1 2 1 2 2 2 1 2 2
F 8: 1 1 2 2 1 2 1 2 1 1 2 2
F 9: 2 1 2 1 2 2 2 1 1 1 1 2
F10: 2 1 1 1 1 2 2 2 1 2 1 2
F11: 1 1 2 1 1 1 2 1 2 2 2 2

Design 11 - 18 Runs, 30 Factors, 1 @ 3 Levels, 29 @ 2 Levels:
F 1: 2 3 3 3 2 1 3 1 1 1 2 2 2 1 1 3 3 2
F 2: 2 1 1 2 1 2 1 1 2 2 2 1 2 2 1 1 2 1
F 3: 1 2 2 1 1 2 1 1 2 1 2 1 2 2 2 2 1 1
F 4: 1 2 2 2 2 2 1 1 1 1 1 2 1 2 2 1 2 1
F 5: 2 1 2 1 1 2 2 1 1 1 2 1 1 2 2 1 2 2
F 6: 1 1 1 2 1 2 2 1 2 2 1 2 1 1 1 2 2 2
F 7: 2 1 2 2 2 1 1 1 1 2 1 2 1 2 1 2 1 2
F 8: 1 1 2 2 1 1 2 1 2 2 1 2 2 2 2 1 1 1
F 9: 2 1 2 1 1 1 2 2 1 1 2 2 1 1 2 2 2 1
F10: 2 1 1 2 2 1 2 1 2 1 1 1 2 1 2 1 2 2
F11: 2 1 2 2 1 2 2 2 1 1 1 2 2 1 1 1 2 1
F12: 2 2 1 1 1 2 1 2 1 2 1 2 2 1 1 2 2 1
F13: 1 2 1 1 2 1 2 1 2 1 1 2 2 2 1 2 2 1
F14: 1 2 1 2 1 1 2 2 1 2 2 1 2 1 2 1 1 2
F15: 1 1 2 2 1 1 1 2 1 2 2 1 2 1 2 2 2 1
F16: 2 1 1 1 1 2 2 1 2 2 1 2 2 1 2 2 1 1
F17: 2 2 2 1 1 1 1 2 2 2 1 1 1 2 1 1 2 2
F18: 1 1 2 1 2 2 1 1 1 2 1 2 2 1 2 1 2 2
F19: 1 1 1 2 1 2 2 2 1 1 2 2 1 2 1 2 1 2
F20: 1 1 2 1 2 1 2 2 2 1 2 2 2 2 1 1 1 1
F21: 1 1 1 2 2 1 1 2 2 2 2 2 1 1 2 1 2 1
F22: 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2
F23: 1 1 1 1 2 2 2 2 2 1 1 1 1 1 2 2 2 2
F24: 1 1 2 2 1 1 1 2 2 1 1 1 2 2 1 2 2 2
F25: 2 2 1 2 1 1 1 2 2 1 1 2 1 2 2 1 1 2
F26: 2 2 1 2 1 1 1 1 2 1 2 2 1 1 2 2 2 1
F27: 1 2 2 1 1 1 2 1 2 2 2 2 1 1 1 1 2 2
F28: 1 2 1 1 1 1 2 2 1 1 1 2 2 2 2 1 2 2
F29: 1 2 2 2 1 2 1 1 2 1 2 2 2 1 1 1 1 2
F30: 2 1 2 1 1 1 1 2 2 1 1 2 2 1 2 2 1 2

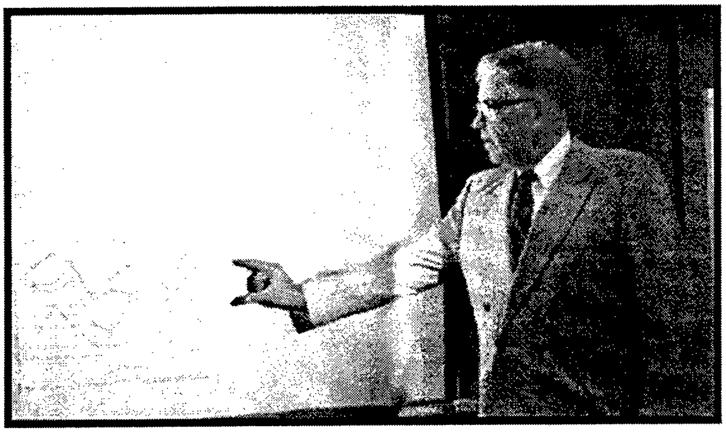
Design 12 - 18 Runs, 24 Factors, 2 @ 3 Levels, 22 @ 2 Levels:
F 1: 2 1 2 3 1 3 1 3 2 1 2 2 3 2 3 1 3 1
F 2: 3 1 1 2 1 2 1 3 2 1 3 2 3 3 2 3 1 2
F 3: 2 1 2 1 2 2 2 1 1 1 2 1 2 1 2 1 2 1
F 4: 1 2 1 2 1 2 2 1 1 1 2 2 2 2 1 1 2 1
F 5: 2 1 1 1 2 2 1 1 2 1 2 2 1 1 1 2 2 2
F 6: 1 2 2 2 2 1 1 1 1 1 2 1 1 2 2 1 2 2
F 7: 2 1 2 1 2 2 1 1 1 2 1 2 1 2 2 1 1 2
F 8: 2 1 2 2 1 2 1 1 1 2 1 1 2 1 1 2 2 2
F 9: 2 1 1 2 1 2 2 2 1 2 2 1 1 1 2 1 1 2
F10: 1 2 2 1 1 2 1 1 2 2 2 1 2 1 2 2 1 1
F11: 1 1 2 2 1 2 1 1 2 2 2 1 1 2 1 1 2 2
F12: 1 1 2 1 1 1 2 2 1 2 2 2 2 1 2 1 1 2
F13: 2 1 1 2 1 1 2 1 2 2 2 1 2 2 1 1 1 2
F14: 1 2 1 1 2 2 1 2 2 2 2 1 2 1 1 1 1 2
F15: 1 1 1 2 2 1 1 2 1 2 2 2 1 1 2 2 2 1
F16: 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2
F17: 1 1 1 1 1 2 2 2 2 1 1 1 1 2 2 2 2 2
F18: 1 1 2 2 2 1 1 2 2 1 1 2 2 1 1 1 2 2
F19: 2 2 1 2 2 1 1 1 1 1 1 1 2 2 2 2 1 2
F20: 1 1 2 2 2 1 2 1 2 1 2 1 2 1 2 2 1 1
F21: 1 1 2 1 2 2 2 2 1 1 1 1 2 2 1 2 1 2
F22: 1 1 2 2 2 2 1 2 1 2 2 1 1 2 1 2 1 1
F23: 1 1 1 2 2 2 2 1 2 2 1 2 2 2 1 1 1 1
F24: 1 2 2 2 1 2 2 1 1 1 2 2 1 1 1 2 1 2

APPENDIX:

CONFERENCE SNAPSHOTS

INTENTIONALLY LEFT BLANK.

The Short Course
Quality Control: Modeling the Deming Paradigm

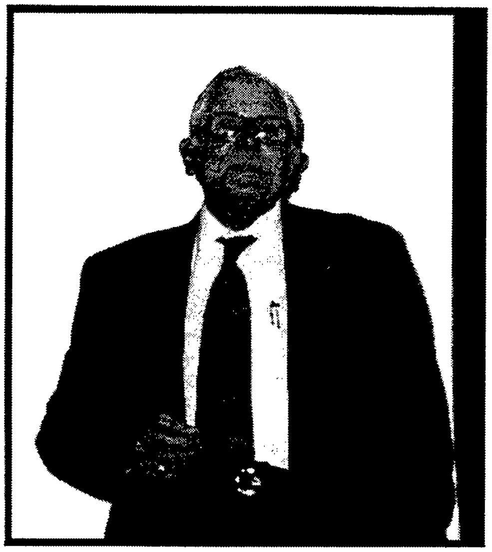


Left: Thompson

Left to Right: Grimes, Webb, Harris, Moss, Prather

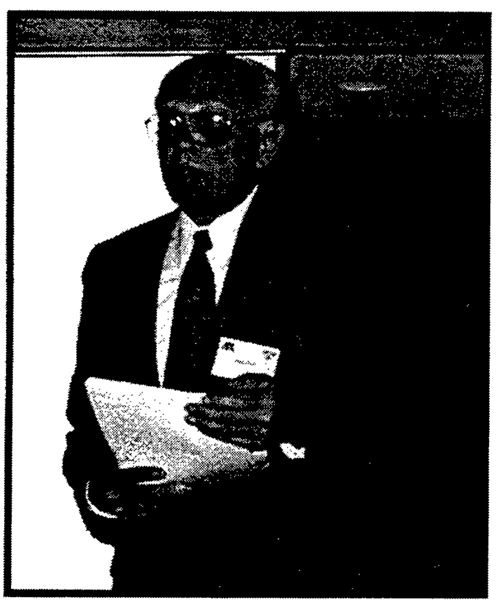
Left to Right: Celmins, Cruess, Burge, Tang

40 Years of Experimentation at Fort Hunter Liggett

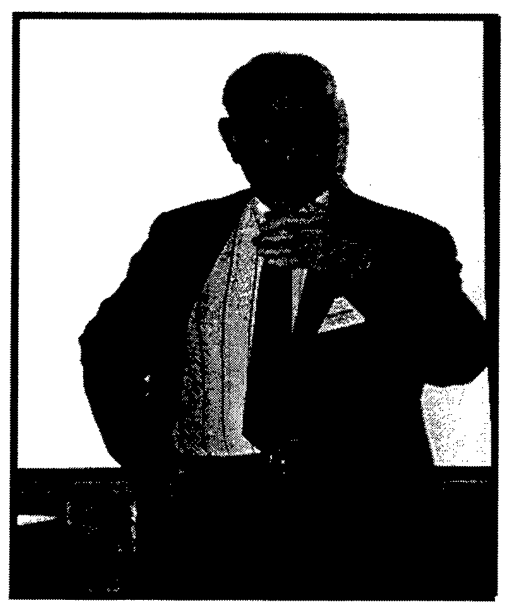


Hollis

Bryson

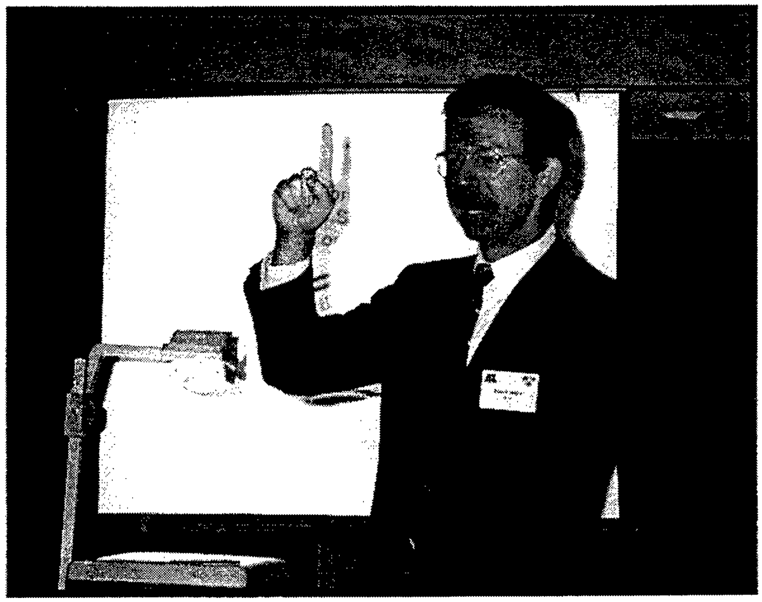


Hill



Alberts

Getting Their Points Across

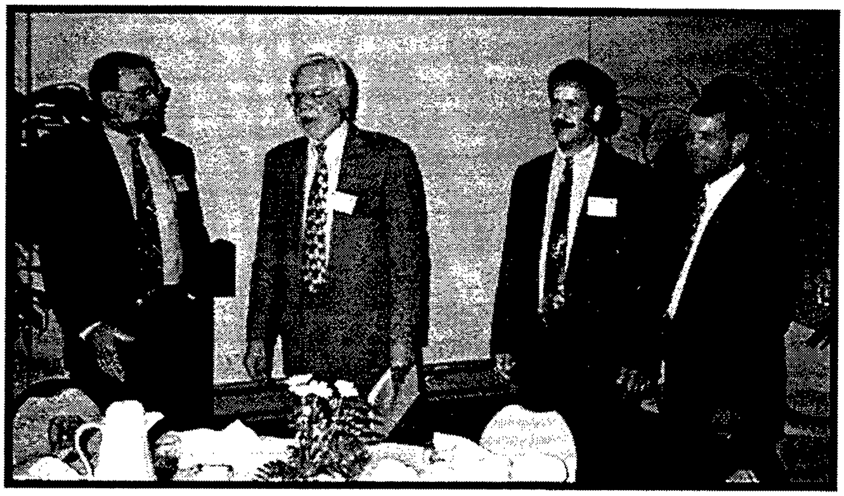


Seglie

Grynovicki

Rob Kass

Leisure Moments



Left to Right: West, Coyle, Bodt, Barr

Left to Right: Launer, Moss, Russell

Left to Right: Dutoit, Grynovicki, Wegman, Cruess, Russell

ATTENDANCE LIST

U.S. ARMY CONFERENCE ON APPLIED STATISTICS
23–25 OCTOBER 1996

John Ackerman
U.S. Army Research Development, and
Engineering Center
ATTN: AMSTA-AR-QAT-P Bldg 62
Picatinny Arsenal, NJ 07806-5000

Henry Alberts
Defense Systems Management RCID
Fort Belvoir, VA 22060

MAJ Suzanne M. Beers
HQ AFOTEC/CNP
8500 Gibson Blvd, SE
Kirtland Air Force Base, NM 87117-5558

Barney Bissinger
NAVICP
281 W. Main Street
Middletown, PA 17057

Barry Bodt
U.S. Army Research Laboratory
ATTN: AMSR-SC-S
Aberdeen Proving Ground, MD 21005-5067

Ann E. M. Brodeen
U.S. Army Research Laboratory
ATTN: AMSRL-IS-CI
Aberdeen Proving Ground, MD 21005-5067

Denise Bullock
Waterways Experiment Station
3909 Halls Ferry Road
Vicksburg, MS 39180-6199

Laura S. Bunch
Waterways Experiment Station
3909 Halls Ferry Road
Vicksburg, MS 39180-6199

J. Robert Burge
Walter Reed Army Institute of Research
Washington, DC 20307-5100

Aivars Celmins
U.S. Army Research Laboratory
Bldg 394
Aberdeen Proving Ground, MD 21005-5067

Alonzo Church, Jr.
Church Associates
136 Hudson St.
Hudson, OH 44236

Joseph Collins
U.S. Army Research Laboratory
ATTN: AMSRL-SL-BV
Aberdeen Proving Ground, MD 21005-5068

David Cruess
USUHS Medical School
4301 Jones Bridge Road
Bethesda, MD 70814-4799

Paul J. Deason
USA TRAC-WSMR
ATTN: ATRC-WAD
White Sands Missile Range, NM
88002-5505

Niki C. Deliman
Waterways Experiment Station
3909 Halls Ferry Road
Vicksburg, MS 39180-6199

Michael Dewitz
U.S. Army Materiel Sys Analysis Activity
ATTN: AMXSY-AP
392 Hopkins Road
Aberdeen Proving Ground, MD 21005-5071

Jack Dowling
Consultant
905 Lighthouse Ave
Pacific Grove, CA 93950

Patrick J. Driscoll
USMA Dept of Math Sciences
West Point, NY 10996-1789

Eugene Dutoit
Dismounted Battlespace Battle Lab
ATTN: ATSH-WCS
Fort Benning, GA 31905-5000

Samuel Frost
EAC
ATTN: CSTE-EAC-AD/ES
4120 Susquehanna Ave
Aberdeen Proving Ground, MD 21005-3013

Bruce W. Gafner
TRAC-WSMR
ATTN: ATRC-WAD
White Sands Missile Range, NM 88002

Duane J. Gotvald
USA MICOM
Huntsville, AL 35898

Fred M. Grimes
OPTEC TEXCOM Close Combat
Fort Hood, TX 76544

Jock O. Grynovicki
U.S. Army Research Laboratory
Human Research and Engr Dir
Bldg 459
Aberdeen Proving Ground, MD 21005-5067

Linda Hall
U.S. Army Aberdeen Test Center
ATTN: STEAC-EN-BA
Aberdeen Proving Ground, MD 21005

Bernard Harris
University of Wisconsin
Dept of Statistics
1210 W. Dayton St
Madison, WI 53706

Floyd I. Hill
6338 Crosswoods Drive
Falls Church, VA 22044

Rick Jernigan
TEXCOM ECSTD CS
Fort Hood, TX 76544

Barbara Kaschenbach
U.S. Army Aberdeen Test Center
ATTN: STEAC-EN-BA
Aberdeen Proving Ground, MD 21005-5069

Rick Kass
Test & Experimentation Command
Fort Hood, TX 76544

Robert Launer
US ARO
P.O. Box 12211
Research Triangle Park, NC 27709-2211

Doug Mackey
TRAC-WSMR
White Sands Missile Range, NM 88002

Bill Meyer
U.S. Army Yuma Proving Ground
ATTN: STRYD-RS
Yuma, AZ 85365

Linda Moss
U.S. Army Research Laboratory
ATTN: AMSRL-SL-BV
Aberdeen Proving Ground, MD 21005-1531

David H. Olwell
USMA Dept. Of Mathematical Sciences
West Point, NY 10996-1786

Buck Ozment
Computer Sciences Corp.
4815 Bradford Dr.
Huntsville, AL 35805

Michael Prather
U.S. Army Evaluation Analysis Center
ATTN: CSTE-EAC-CS
Bldg 314
Aberdeen Proving Ground, MD 21005-5055

C. R. Rao
Penn State University
Stat Dept Thomas Bldg
University Park, PA 16802

Nancy A. Renfroe
Waterways Experiment Station
3909 Halls Ferry Road
Vicksburg, MS 39180-6199

Carl T. Russell
TEXCOM Experimentation Center
ATTN: CSTE-TEC-CC
Fort Hunter Liggett, CA 93928-8000

Richard Saunders
Scientific Support Lab TEC
P.O. Box 918
Jolon, CA 93940

Jayaram Sethuraman
Florida State Univ, Statistics Dept
Tallahassee, FL 32306

Douglas B. Tang
Walter Reed Army Institute of Research
Washington, DC 20012

Deloris Testerman
Yuma Proving Ground
ATTN: STEYP-RS (VPG)
Yuma, AZ 85365

James R. Thompson
Rice University
Dept of Statistics
Houston, TX 77251-1892

Thomas R. Walker
U.S. Army Aberdeen Test Center
ATTN: STEAC-EN-AA
Aberdeen Proving Ground, MD 21005

R. John Weaver
Pharmacia & Upjohn
16007 Prairie Road
Schoolcraft, MI 49087-9739

David W. Webb
U.S. Army Research Laboratory
ATTN: AMSRL-WM-PB
Aberdeen Proving Ground, MD 21005-5066

Professor Edward J. Wegman
George Mason University
MS 4A7, 4400 University Drive
Fairfax, VA 22030-4444

W. Max Woods
Naval Postgraduate School
Monterey, CA 93943

Brian Barr
HQ, Test and Experimentation Command
ATTN: CSTE-TTD
Bldg 9102
Fort Hood, TX 76544-5065

Grail Brookshire
HQ, Test and Experimentation Command
ATTN: TESCO
Bldg 9102
Fort Hood, TX 76544-5065

Edward C. Buntz
2527 Smith Road
Bradley, CA 93246

Professor Herman Chernoff
Dept. of Statistics SC713
Harvard University
Cambridge, MA 02138-2901

Professor Jay Conover
College of Business Administration
Texas Tech University
Lubbock, TX 79409

Honorable Philip E. Coyle III
OSD/DOT&E
1700 Defense Pentagon
Washington, DC 20301-1700

Dr. Donald Gaver
Dept. of Operations Research
Naval Postgraduate School
Monterey, CA 93943

Professor Ulf Grenander
Brown University
Box F.
Providence, RI 02906

Mr. Walt Hollis
DA/DUSA(OR)
The Pentagon
Washington, DC 20310

Professor Rob Kass
Dept. of Statistics
Carnegie-Mellon University
Pittsburgh, PA 15213-3890

Kragg Kysor
U.S. Army Research Laboratory
ATTN: AMSRL-HR-S
Aberdeen Proving Ground, MD 21005

COL James R. Prouty
TEXCOM Experimentation Center
Fort Hunter Liggett, CA 93928-8000

Dr. Ernest Seglie
OSD/DOT&E
1700 Defense Pentagon
Washington, DC 20301-1700

John Schuler
Scientific Support Laboratory
Fort Hunter Liggett, CA 93928-8000

Mike Tedeschi
8813 Deer Trail Road
Bradley, CA 93426

Bill West
25230 Baronet Drive
Salinas, CA 93908

Dennis Whitmer
Scientific Support Laboratory
Fort Hunter Liggett, CA 93928-8000

Jovencia Williams
18865 Moro Circle
Prunedale, CA 93907

Frank Apicella
HQ U.S. Army Operational Evaluation Cmd
ATTN CSTE-ECC
4501 Ford Avenue
Alexandria, VA 22302-1458

Robin Woo
P.O. Box 490
Marina, CA 93933

Cecil Bone
Rt. 5, Box 5350A E. Lakeshore Dr.
Belton, TX 96513

Addresses unknown for the following:
Larry Evens
Pete Hedges
Jack Holbrook
Gary Love

INTENTIONALLY LEFT BLANK.

NO. OF COPIES	ORGANIZATION
2	DEFENSE TECHNICAL INFORMATION CENTER DTIC DDA 8725 JOHN J KINGMAN RD STE 0944 FT BELVOIR VA 22060-6218
1	HQDA DAMO FDQ DENNIS SCHMIDT 400 ARMY PENTAGON WASHINGTON DC 20310-0460
1	CECOM SP & TRRSTRL COMMCTN DIV AMSEL RD ST MC M H SOICHER FT MONMOUTH NJ 07703-5203
1	PRIN DPTY FOR TCHNLGY HQ US ARMY MATCOM AMCDCG T M FISETTE 5001 EISENHOWER AVE ALEXANDRIA VA 22333-0001
1	PRIN DPTY FOR ACQUSTN HQS US ARMY MATCOM AMCDCG A D ADAMS 5001 EISENHOWER AVE ALEXANDRIA VA 22333-0001
1	DPTY CG FOR RDE HQS US ARMY MATCOM AMCRD BG BEAUCHAMP 5001 EISENHOWER AVE ALEXANDRIA VA 22333-0001
1	ASST DPTY CG FOR RDE HQS US ARMY MATCOM AMCRD COL S MANESS 5001 EISENHOWER AVE ALEXANDRIA VA 22333-0001

NO. OF COPIES	ORGANIZATION
1	DPTY ASSIST SCY FOR R&T SARD TT F MILTON THE PENTAGON RM 3E479 WASHINGTON DC 20310-0103
1	DPTY ASSIST SCY FOR R&T SARD TT D CHAIT THE PENTAGON WASHINGTON DC 20310-0103
1	DPTY ASSIST SCY FOR R&T SARD TT K KOMINOS THE PENTAGON WASHINGTON DC 20310-0103
1	DPTY ASSIST SCY FOR R&T SARD TT B REISMAN THE PENTAGON WASHINGTON DC 20310-0103
1	DPTY ASSIST SCY FOR R&T SARD TT T KILLION THE PENTAGON WASHINGTON DC 20310-0103
1	OSD OUSD(A&T)/ODDDR&E(R) J LUPO THE PENTAGON WASHINGTON DC 20301-7100
1	INST FOR ADVNCD TCHNLGY THE UNIV OF TEXAS AT AUSTIN PO BOX 202797 AUSTIN TX 78720-2797
1	DUSD SPACE 1E765 J G MCNEFF 3900 DEFENSE PENTAGON WASHINGTON DC 20301-3900
1	USAASA MOAS AI W PARRON 9325 GUNSTON RD STE N319 FT BELVOIR VA 22060-5582

NO. OF COPIES	ORGANIZATION
1	CECOM PM GPS COL S YOUNG FT MONMOUTH NJ 07703
1	GPS JOINT PROG OFC DIR COL J CLAY 2435 VELA WAY STE 1613 LOS ANGELES AFB CA 90245-5500
1	ELECTRONIC SYS DIV DIR CECOM RDEC J NIEMELA FT MONMOUTH NJ 07703
3	DARPA L STOTTS J PENNELLA B KASPAR 3701 N FAIRFAX DR ARLINGTON VA 22203-1714
1	SPCL ASST TO WING CMNDR 50SW/CCX CAPT P H BERNSTEIN 300 O'MALLEY AVE STE 20 FALCON AFB CO 80912-3020
1	USAF SMC/CED DMA/JPO M ISON 2435 VELA WAY STE 1613 LOS ANGELES AFB CA 90245-5500
1	US MILITARY ACADEMY MATH SCI CTR OF EXCELLENCE DEPT OF MATHEMATICAL SCI MDN A MAJ DON ENGEN THAYER HALL WEST POINT NY 10996-1786
1	DIRECTOR US ARMY RESEARCH LAB AMSRL CS AL TP 2800 POWDER MILL RD ADELPHI MD 20783-1145

NO. OF COPIES	ORGANIZATION
1	DIRECTOR US ARMY RESEARCH LAB AMSRL CS AL TA 2800 POWDER MILL RD ADELPHI MD 20783-1145
3	DIRECTOR US ARMY RESEARCH LAB AMSRL CI LL 2800 POWDER MILL RD ADELPHI MD 20783-1145
	ABERDEEN PROVING GROUND
2	DIR USARL AMSRL CI LP (305)

NO. OF COPIES	ORGANIZATION
1	OSD/DOT&E ATTN HONORABLE P E COYLE III 1700 DEFENSE PENTAGON WASHINGTON DC 20301-1700
1	DA/DUSA(OR) ATTN W HOLLIS THE PENTAGON WASHINGTON DC 20310
1	OSD/DOT&E ATTN DR E SEGLIE 1700 DEFENSE PENTAGON WASHINGTON DC 20301-1700
1	US ARMY RESEARCH DEVELOPMENT AND ENGINEERING CENTER ATTN AMSTA AR QAT P J ACKERMAN BLDG 62 PICATINNY ARSENAL NJ 07806-5000
1	DEFENSE SYS MANAGEMENT RCID ATTN H ALBERTS FORT BELVOIR VA 22060
5	WALTER REED ARMY INSTITUTE OF RESEARCH ATTN J R BURGE WASHINGTON DC 20307-5100
5	WALTER REED ARMY INSTITUTE OF RESEARCH ATTN D B TANG WASHINGTON DC 20012
7	USA TRAC-WSMR ATTN ATRC WAD P J DEASON (5 CPS) B W GAFNER D MACKEY WHITE SANDS MISSILE RANGE NM 88002-5505
1	USMA DEPT OF MATH SCIENCES ATTN P J DRISCOLL WEST POINT NY 10996-1789
1	USA MICOM ATTN D J GOTVALD HUNTSVILLE AL

NO. OF COPIES	ORGANIZATION
1	OPTEC TEXCOM CLOSE COMBAT ATTN F M GRIMES FORT HOOD TX 76544
4	WATERWAYS EXPERIMENT STATION ATTN D BULLOCK L S BUNCH NIKI C DELIMAN N A RENFROE 3909 HALLS FERRY ROAD VICKSBURG MS 39180-6199
1	TEXCOM ECSTD CS ATTN R JERNIGAN FORT HOOD TX 76544
1	TEST AND EXPERIMENTATION CMD ATTN R KASS FORT HOOD TX 76544
5	USARO ATTN R LAUNER PO BOX 12211 RESEARCH TRIANGLE PARK NC 27709-2211
6	US ARMY YUMA PROVING GROUND ATTN STRYD RS B MEYERS STEYP RS (VPG) D TESTERMAN (5 CPS) YUMA AZ 85365
5	DISMOUNTED BATTLESPACE BATTLE LAB ATTN ATSH WCS E DUTOIT FORT BENNING GA 31905-5000
5	JNTF/SE ATTN DR RUSSELL 730 IRWIN AVE FALCON AIR FORCE BASE CO 80912-7300
1	NAVICP ATTN B BISSINGER 281 W MAIN STREET MIDDLETOWN PA 17057
5	USUHS MEDICAL SCHOOL ATTN D CRUESS 4301 JONES BRIDGE ROAD BETHESDA MD 70814-4799

NO. OF COPIES	ORGANIZATION
1	NAVAL POSTGRADUATE SCHOOL ATTN W M WOODS MONTEREY CA 93943
1	NAVAL POSTGRADUATE SCHOOL ATTN DR D GAVER DEPT OF OPERATIONS RESEARCH MONTEREY CA 93943
1	USMA DEPT OF MATH SCIENCES ATTN D H OLWELL WEST POINT NY 10996-1786
1	HQ AFOTEC/CNP ATTN MAJ S M BEERS 8500 GIBSON BLVD SE KIRTLAND AIR FORCE BASE NM 87117-5558
1	SCIENTIFIC SUPPORT LAB TEC ATTN R SAUNDERS PO BOX 918 JOLON CA 93940
1	UNIVERSITY OF WISCONSIN ATTN B HARRIS DEPT OF STATISTICS 1210 W DAYTON ST MADISON WI 53706
1	PENN STATE UNIVERSITY ATTN C R RAO STAT DEPT THOMAS BLDG UNIVERSITY PARK PA 16802
1	FLORIDA STATE UNIV ATTN J SETHURAMAN STATISTICS DEPT TALLAHASSEE FL 32306
1	RICE UNIVERSITY ATTN J R THOMPSON DEPT OF STATISTICS HOUSTON TX 77251-1892
1	GEORGE MASON UNIVERSITY ATTN PROFESSOR E J WEGMAN MS 4A7 4400 UNIVERSITY DRIVE FAIRFAX VA 22030-4444

NO. OF COPIES	ORGANIZATION
1	HARVARD UNIVERSITY ATTN PROFESSOR H CHERNOFF DEPT OF STATISTICS SC713 CAMBRIDGE MA 02138-2901
1	COLLEGE OF BUSINESS ADMIN ATTN PROFESSOR J CONOVER TEXAS TECH UNIVERSITY LUBBOCK TX 79409
1	BROWN UNIVERSITY ATTN PROFESSOR U GRENANDER BOX F PROVIDENCE RI 02906
1	CARNEGIE-MELLON UNIVERSITY ATTN PROFESSOR ROB KASS DEPT OF STATISTICS PITTSBURGH PA 15213-3890
1	CHURCH ASSOCIATES ATTN A CHURCH JR 136 HUDSON ST HUDSON OH
1	J DOWLING 905 LIGHTHOUSE AVE PACIFIC GROVE CA 93950
1	F I HILL 6338 CROSSWOODS DRIVE FALLS CHURCH VA 22044
1	COMPUTER SCIENCES CORP ATTN B OZMENT 4815 BRADFORD DR HUNTSVILLE AL 35805
1	PHARMACIA & UPJOHN ATTN R J WEAVER 16007 PRAIRIE ROAD SCHOOLCRAFT MI
5	NIST STATISTICAL ENGINEERING DIV ATTN DR M VANGEL BLDG 101 A337 GAITHERSBURG MD 20899-0001

NO. OF COPIES	ORGANIZATION
	ABERDEEN PROVING GROUND, MD
1	DIR USAMSAA ATTN AMXSY AP M DEWITZ
5	DIR USAATC ATTN STEAC EN AA T R WALKER STEAC EN BA L HALL B KASCHENBACH L HALL B KASCHENBACH
1	DIR USAEAC ATTN CSTE EAC CS M PRATHER BLDG 314
1	DIR EAC ATTN CSTE EAC AD/ES S FROST 4120 SUSQUEHANNA AVE
21	DIR USARL ATTN AMSRL CI CA AIVARS CELMINS AMSRL HR S K KYSOR J O GRYNOVICKI (5 CPS) AMSRL SC S B BODT (5 CPS) AMSRL IS CI A E M BRODEEN AMSRL SL BV L MOSS JOSEPH COLLINS AMSRL WM PB D W WEBB AMSRL IS M TAYLOR (5 CPS)

INTENTIONALLY LEFT BLANK.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
July 1997

3. REPORT TYPE AND DATES COVERED
Final, 23-25 October 1996

4. TITLE AND SUBTITLE
Proceedings of the Second Annual U.S. Army Conference on Applied Statistics, 23-25 October 1996

5. FUNDING NUMBERS
AH80

6. AUTHOR(S)
Barry A. Bodt, Editor

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: AMSRL-IS-CI
Aberdeen Proving Ground, MD 21005-5067

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-SR-59

9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The second U.S. Army Conference on Applied Statistics was held 23-25 October 1996 at the Monterey Beach Hotel, Monterey, CA, and hosted by the TEXCOM Experimentation Center at nearby Fort Hunter Liggett. The conference was cosponsored by the U.S. Army Research Laboratory; the U.S. Army Research Office; the U.S. Military Academy; the U.S. Army Training and Doctrine Command Analysis Center, White Sands Missile Range; the Walter Reed Army Institute of Research; and the National Institute for Standards and Technology. Papers given at the conference addressed the development of new statistical techniques, application of existing methodologies to Army problems, and panel discussion of statistical challenges in an Army setting. A special session was included to commemorate Fort Hunter Liggett, the dedicated civilians and military who have worked there, and the countless contributions to Army testing that were developed and practiced there. This document is a compilation of available papers offered at the conference.

14. SUBJECT TERMS
applied statistics, experimental design, statistical inference

15. NUMBER OF PAGES
206

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18 298-102

INTENTIONALLY LEFT BLANK.

USER EVALUATION SHEET/CHANGE OF ADDRESS

This Laboratory undertakes a continuing effort to improve the quality of the reports it publishes. Your comments/answers to the items/questions below will aid us in our efforts.

1. ARL Report Number/Author ARL-SR-59 (Bodt, Editor) ____________ Date of Report July 1997 ________

2. Date Report Received ______________________________

3. Does this report satisfy a need? (Comment on purpose, related project, or other area of interest for which the report will be used.) ______________________________

4. Specifically, how is the report being used? (Information source, design data, procedure, source of ideas, etc.) ______________________________

5. Has the information in this report led to any quantitative savings as far as man-hours or dollars saved, operating costs avoided, or efficiencies achieved, etc? If so, please elaborate. ______________________________

6. General Comments. What do you think should be changed to improve future reports? (Indicate changes to organization, technical content, format, etc.) ______________________________

CURRENT ADDRESS

Organization

Name E-mail Name

Street or P.O. Box No.

City, State, Zip Code

7. If indicating a Change of Address or Address Correction, please provide the Current or Correct address above and the Old or Incorrect address below.

OLD ADDRESS

Organization

Name

Street or P.O. Box No.

City, State, Zip Code

(Remove this sheet, fold as indicated, tape closed, and mail.)

(DO NOT STAPLE)